ON STOCK RATIONING POLICIES FOR CONTINUOUS REVIEW INVENTORY SYSTEMS

A THESIS
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCES
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By

Önder Bulut

July 2005

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. M. Murat Fadıloğlu (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Nesim Erkip

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Osman Alp

Approved for the Institute of Engineering and Sciences:

Prof. Mehmet Baray
Director of Institute of Engineering and Sciences

ABSTRACT

ON STOCK RATIONING POLICIES FOR CONTINUOUS REVIEW INVENTORY SYSTEMS

Önder Bulut

M.S. in Industrial Engineering

Supervisor: Asst. Prof. M. Murat Fadıloğlu

July 2005

Rationing is an inventory policy that allows prioritization of different demand classes. In this thesis, we analyze the stock rationing policies for continuous review systems. We clarify some of the ambiguities present in the current literature. Then, we propose a new method for the exact analysis of lot-for-lot inventory systems with backorders under rationing policy. We show that if such an inventory system is sampled at multiples of supply leadtime, the state of the system evolves according to a Markov chain. We provide a recursive procedure to generate the transition probabilities of the embedded Markov chain. It is possible to obtain the steady-state probabilities of interest with desired accuracy by considering a truncated version of the chain. Finally, we propose a dynamic rationing policy, which makes use of the information on the status of the outstanding replenishment orders. We conduct a simulation study to evaluate the performance of the proposed policy.

Keywords: Stochastic inventory models, stock rationing, multiple demand classes, embedded Markov chains, solution of infinite state space Markov chains

ÖZET

SÜREKLİ GÖZDEN GEÇİRİLEN ENVANTER SİSTEMLERİNDE STOK TAYINLAMA POLİTİKALARI ÜZERİNE

Önder Bulut
Endüstri Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Yrd. Doç. M. Murat Fadıloğlu
Temmuz 2005

Stok tayınlama politikaları, farklı talep sınıfları için bir tür öncelik mekanizması oluşturulmasına yarar. Bu tez çalışmasında sürekli gözden geçirilen envanter sistemleri için stok tayınlama politikaları incelenmiştir. Konuyla ilgili şu ana kadar yapılmış önemli çalışmalardaki muğlaklıklar giderildikten sonra ardısmarlamalı bire-bir envanter sisteminde kritik seviye stok tayınlama politikası için kesin analize imkan veren yeni bir yöntem önerilmektedir. Bu yöntemle gömülü bir Markov zinciri tanımlanmakta ve bu zincirinin geçiş olasılıkları bir özyineli prosedürle üretilmektedir. Kalıcı durum olasılıklarının Markov zincirinin bir bölümü kullanılarak istenilen doğrulukta bulunabileceği gösterilmiştir. Son olarak, beklenen siparişlerin ulaşmalarına kalan zaman bilgisini kullanan dinamik bir stok tayınlama politikası önerilmektedir. Önerilen politikanın performans değerlendirilmesi benzetim deneyleri kullanılarak yapılmıştır.

Anahtar sözcükler: Rassal envanter modelleri, stok tayınlama, çoklu talep sınıfları, gömülü Markov zinciri, sonsuz durum uzaylı Markov zincirlerinin çözümü

ACKOWLEDGEMENT

I would like to express my sincere gratitude to my advisor Asst. Prof. M. Murat Fadıloğlu for all the trust and encouragement.

I am indebted to Prof. Nesim Erkip and Asst. Prof. Osman Alp for accepting to read and review this thesis and for their invaluable suggestions

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

For inventory systems that experience classes of demand for a single item, stock rationing is a well-known tool to differentiate customers. More specifically, stock rationing is an inventory policy that allows prioritization of different demand classes in order to provide different levels of service and to achieve higher operational efficiency. It is possible to maintain high service levels for certain demand classes while keeping inventory costs at bay by providing lower service levels to certain other demand classes. Demand classes are categorized on the basis of their shortage costs. The highest priority class is the one with the largest shortage costs, and the lowest priority has the smallest shortage costs. If there are *n* demand classes then class 1 has the highest priority, class n has the lowest. In an inventory system with backordering, the ordering of the unit backordering costs is $\pi_1 > \pi_2 > ... > \pi_n$ and similarly the ordering of the time dependent backordering costs is $\hat{\pi}_1 > \hat{\pi}_2 > ... > \hat{\pi}_n$.

The systems that have multiple customer classes generating demand for a unit product are frequently observed in real life. For example, in spare parts inventory management, system can experience urgent and critical orders in case of breakdowns that have high shortage costs. On the other hand, the orders due to the planned maintenance activities may be less critical and

usually have less shortage costs. Again for the spare parts inventory system with multiple end products, same part can be used in many end products that have different importance and criticality. Thus, the demand for any spare part from these end products should be prioritized.

Another example is a two-echelon inventory system consisting of a warehouse and many retailers. In case of stockouts, retailers may place urgent, more critical orders to the warehouse. Again in this setting, if the retailers are located on the basis of regional characteristics of the market area, this situation possibly implies different demand classes and the prioritization of retail orders are beneficial. As another example, in multi-echelon systems the inventory locations in the same echelon can allow shipments between themselves in order to increase the service levels for their direct customers. However, for any inventory location, direct customer orders have superiority over those intershipment orders that are placed by other locations.

Rationing is also a well-known tool in service sectors for customer differentiation. Hotel or airline companies ration their limited capacity according to the priorities of their different customer classes. In this setting, in addition to the rationing decision the key concern is deciding the prices charged to each demand classes. Since the capacity is fixed, in these problems the decisions of when to order and how much to order are irrelevant.

For the systems with multiple demand classes, the most common and easiest strategies are managing individual, separate stock systems for each customer class and managing a common stock pool to serve all the classes without any differentiation. The separate stock strategy permits to assign

different service levels to each customer class, but the positive effect of risk pooling is disregarded. The variability of the demand is higher in this strategy and therefore the whole system has to hold more safety stock to guarantee the desired service levels. On the other hand, common stock strategy uses the pooling effect. Yet this policy causes unnecessary inventory investments for the lower priority classes, because the system provides the highest service level required by the higher priority classes to all demand classes. Inventory rationing policies capture the pooling effect of the demand and in addition to this, they have the flexibility of providing different service levels to different customer classes.

It is possible to define many different kinds of rationing policies, but the mechanism through which any rationing policy is implemented is to stop serving a lower priority class when the inventory on hand inventory drops below a certain threshold level. Under this level only higher priority classes are served and this results in higher service levels for these classes. If there are more than two demand classes, then there is more than a single threshold level. The threshold levels may change dynamically according to the number and the ages of outstanding orders or static threshold levels may be used. The rationing policy with static threshold levels is known as the critical level policy in the literature. For the classic *(r, Q)* system with critical level rationing, the policy can be defined by the decision variables $(\vec{K}, r, Q)$ where $\vec{K}$ represents the vector which consists of (*n-1)* critical levels. The critical level for the highest priority class is not specified since it is always set to 0. If $n = 2$ then the policy is (K, r, Q).

In the literature, there is no result characterizing the optimum policy

structure for the stock rationing problem except for the capacitated make-to-stock production systems. For exponential leadtime Ha (1997a) characterize the optimum policy for lost sales case and Ha (1997b) does for the backordering case. Ha (2000) describes the optimum policy for Erlang distributed leadtime and lost sales case, later Gayon et al. (2005) partially define the optimum policy for the backordering case.

In this thesis, we are interested in traditional inventory systems, which have uncapacitated supply channels. It is obvious that dynamic policies that adjust the critical levels continuously in time by utilizing the information on the number of outstanding orders and their remaining times to arrive, are closer to the unknown optimum structure than the (static) critical level policy. To motivate this fact let us consider an inventory system in which there is an outstanding order that is about to arrive. If the probability of stockout for the higher priority classes in the remaining order arrival time is negligible, then we should choose to satisfy a demand of lower priority class even if the inventory on hand is below the critical level. On the other hand, dynamic decisions considerably complicate the performance evaluation and the optimization of policy parameters. Thus the common practice is to use static threshold levels, i.e. employing critical level policy.

In a backordering environment, stock rationing introduces the problem of making allocation decision of incoming replenishment orders between increasing the stock level and clearing the backorders of different customer classes. The rule that governs this allocation decision is called the clearing mechanism. Without specifying the clearing mechanism, the rationing policy cannot be fully defined for inventory systems with backordering. Consider a system with two demand classes, class1 and class 2, when a replenishment order arrives it is optimal first to clear class 1 backorders due to high

backorder costs of this class. After clearing all class 1 backorders one can give the priority to increasing the stock level up to the critical level. Then if all the order quantity has not been used, s/he can clear the backorders of class 2 starting from the oldest one until using all the remaining order quantity or until the class 2 backorders are depleted. If any units left from the order quantity after all class 2 backorders are filled, those units are added to the inventory to increase the stock level. This clearing mechanism is called *priority clearing* in the literature. Under priority clearing, inventory level cannot exceed K before clearing all backorders. One possible alternative to priority clearing is, after clearing class 1 backorders one can fulfill the backordered demands of class 2 then increase the stock level. One can come up with many other clearing mechanisms. Some different kinds of clearing mechanisms have already been suggested in the literature mostly due to the fact that exact analysis of stock rationing problem with priority clearing is not available. In a lost sales environment there is no clearing concept.

In this thesis, following a general review of stock rationing literature, we present a detailed analysis of the critical level policy based on our observations in parallel with some notes on the main works in the literature that considered critical level policy and different clearing mechanisms. Then we present our contributions to the literature. The setting we consider is a continuously reviewed single location, and single product inventory system with backordering under rationing. We assume deterministic leadtime, L, and two customer classes that generate Poisson demand arrivals with rates λ_1 and λ_2. We first introduce an Embedded Markov Chain approach to the analysis of *(S-1, S)* policy under rationing with the priority clearing mechanism. With this approach we are able to obtain the steady state probabilities of the system with desired accuracy by considering a truncated version of an infinite state space Markov chain. These state probabilities

permit the computation of any long-run performance measure of interest for the system. Finally for the *(r, Q)* policy, we present a dynamic rationing policy that utilizes the information of the number of outstanding orders and their ages. We conduct a simulation study to quantify the gains through this dynamic policy.

We organized the thesis in six chapters. In the following chapter we discuss the related literature, then in Chapter 3 we present our observations on critical level policy and on some of the important works in the literature. In Chapter 4 we present our embedded Markov chain approach for the *(S-1, S)* policy under rationing and in Chapter 5 we introduce a dynamic rationing policy. Finally, we provide an overall summary of the study and address future research directions in Chapter 6.

Chapter 2

LITERATURE REVIEW

Similar to other stochastic inventory problems, stock rationing literature can be categorized based on the review policy (continuous/periodic) and on the consequence of shortages (backorders/lost sales). However, in addition to this general categorization research on stock rationing is also classified according to the assumed rationing policy and by the clearing mechanism for the backorders that defines how to handle the arriving replenishment orders. There is also a parallel literature on the production environment, which effectively considers a capacitated replenishment channel.

Rationing models deal with the prioritization of different demand classes and the initial research on this topic were at 1960s, which generally considered the periodic review systems. Veinott (1965) is the first who analyze a periodic review setting with zero leadtime and backordering. He introduces the concept of critical levels as a rationing policy for multiple demand classes. Topkis (1968) worked on the same model and proved the optimality of time remembering critical level policy for both lost sales and backordering cases. His policy is based on dividing every review period into a finite number of subperiods. By using dynamic programming, he finds a critical level for each subperiod and for each demand class that depends on the time to the next review. Critical level is decreasing with the remaining time to review. For the lost sales case Evans (1968) and for the backordering

case Kaplan (1969) derive essentially the same results for two demand classes.

Nahmias and Demmy (1981) consider a stationary, fixed critical level policy for two demand classes and derive the expected number of backorders for both types of the customer classes for a single period problem and extended the results to an infinite horizon multiperiod problem with the assumed (*s, S*) policy and zero lead time. They assume that demand is realized at the end of each period. Moon and Knag (1998) generalize the work of Nahmias and Demmy (1981) by considering multiple critical levels and by presenting a simulation analyses.

Cohen et al. (1989) consider a periodic review (*s, S*) policy with lost sales and two demand classes. At the end of each period, after the realization of demands, they use the on hand stock to meet the demands of customer classes in the order of priorities.

Frank et al. (2003) analyze a periodic review model with two demand classes, one is stochastic and the other is deterministic. High priority class is the one with deterministic demand. Any unsatisfied stochastic demand is lost. They characterize the complex structure of the optimal policy and propose a much simpler critical level policy for (*s, S*) type replenishment. They assume that the orders arrive instantaneously and aim to use the rationing to gain from fixed ordering cost instead of saving stock for future deterministic demand.

Atkins and Katircioglu (1995) consider service level requirements for each demand classes and propose a heuristic rationing policy that is hard to implement. They have a periodic review setting with fixed lead time and backordering.

CHAPTER 2 LITERATURE REVIEW

In continuous review setting, multiple demand classes and rationing first analyzed by Nahmias and Demmy (1981). Under a (r,Q) policy they consider a setting with unit Poisson arrivals, two demand classes, constant leadtime, and full backordering. They assume a critical level rationing policy and at most one outstanding replenishment order. Instead of finding the optimum policy parameters, i.e.(K,r,Q) and K stands for the critical level, they focus on deriving the expected number of backorders and fill rates for any given parameter set for both demand classes. Moon and Kang (1998) extend the model of Nahmias and Demmy (1981) by considering compound Poisson demand. They analyze the system with a simulation model.

Deshpande et al. (2003) work on exactly the same problem that Nahmias and Demmy (1981) analyze with the exception that they do not have any restriction on the number of outstanding orders. However, in order to get the analytical results of this more general critical level rationing model they do not use the priority clearing mechanism. To derive the steady state inventory level probabilities and to get the operating characteristics of the system they introduce the threshold clearing mechanism that allows clearing low priority backorders before clearing all class 1 backorders and raising the inventory above the critical level. They provide an algorithm to obtain the optimal policy parameters, which result in the minimum expected cost rate. They compare the performance of the threshold clearing with the performance of priority clearing. For the priority clearing, they obtain the optimum levels via simulation.

Melchiors et al. (2000) analyze the model of Nahmias and Demmy (1981) for the lost sales case by preserving the assumption of at most one outstanding order. They allow the critical level to be above the reorder level

and observe some situations that this case is optimal. Melchiors (2003) extends this model by considering multiple Poisson demand classes and they propose the restricted time remembering policy. He divides the constant leadtime in subintervals and by considering the remaining lead time of the outstanding order finds the critical levels which are restricted to be constant over the subintervals.

Teunter and Haneveld (1996) also consider a time remembering policy in the continuous review setting for two demand classes and backordering. They determine the set of remaining lead time values (L_1, L_2...) which imply to reserve 0,1,2… units of stock for high priority customers, i.e. if the remaining time is less than L_1 they do not ration the stock, if it is between L_1 and (L_1+L_2) one item is reserved for the high priority class and so on. They showed that this policy outperforms the critical level policy.

Dekker et al. (1998) work on a spare parts stocking environment with two demand classes and consider (*S-1, S*) inventory policy that allows backordering. They state that they do not have any restriction on the number of outstanding orders. For the critical level rationing, without assuming any clearing mechanism they derive the exact fill rate expression for the non-critical demand class and make an approximation for the critical class fill rate by conditioning on the time that stock level hits the critical level. Afterwards they consider three different clearing mechanisms including the priority clearing and test their approximation under each of these mechanisms using simulation. They also present another approximation for the service level of the critical class, which accounts the effect of the way that the incoming orders are handled.

Similar to Dekker et al. (1998), Dekker et al. (2002) consider the (*S-1, S*)

policy and critical level rationing. However, they assume generally distributed leadtime, multiple demand classes and lost sales. In lost sales case, there is no discussion about clearing. Using queueing results they derive the state probabilities and operating characteristics of the system. They introduce a numerical solution method for optimization.

Ha (1997a) considers a make-to-stock production system with a single production facility, zero setup cost, multiple demand classes and lost sales. He assumes exponentially distributed production leadtime. His system is a capacitated one and order crossing is not possible. Using a queueing model he shows that lot-for-lot policy is optimal for production decision and the critical level policy is optimal for stock rationing decision. Intuitively, for a memoryless system the elapsed time does not provide any information for the arrival time of the replenishment order. Ha (1997b) analyzes the same setting but he allows backordering. He defines the optimal control policy as a monotone switching curve which says that production decision is based on a based stock policy and rationing decision determined by critical level policy which is decreasing in the number of backorders of the non-critical class. Vericourt (2002) considers the multiple demand class extension of Ha (1997b).

Ha (2000) extends Ha (1997a) to an Erlangian production times. He defines the work storage level concept that keeps track of the number of completed Erlang stages for the items in the system. He proves that a critical work storage level is optimal for both production and stock rationing. With a numerical analysis he shows that the critical level policy performs well for the same setting.

Gayon et al. (2005) considers the setting of Ha (2000) but they allow

backordering. Using the work storage level concept, they partially characterize the optimal policy. In addition, when they assume a salvage market without a backorder cost they fully characterize the optimal stock rationing policy.

Kocaga (2004) works on the spare parts service system of a leading semiconductor equipment manufacturer. He considers the same setting of Dekker et al. (1998). However, the non-critical orders allow a fixed demand leadtime to be fulfilled. After deriving the service level expressions for both classes, with a numerical study he shows that significant savings are possible through incorporation of demand leadtimes and rationing.

Arslan et al. (2005) considers a continuous review (r,Q) policy with multiple demand classes, unit Poisson demands and deterministic leadtime. They assume the critical level rationing and analyze this single location inventory system by constructing an equivalent multi stage serial system. The stages in the serial system are defined as inventory systems that face the external demand of corresponding customer class of the original problem. Each stage also sees internal demands from the lower level stages. By assuming a clearing mechanism that clears the backorders at each stage in the order of occurrence, they derive the state probabilities of the system. However, this clearing mechanism allocates the replenishment quantity fairly between the reserve stock for higher classes and the backorders of lower level classes. Thus, it deviates from the priority clearing. They also provide a heuristic for the optimization and describe how their model can be extended to a multi echelon setting.

Zhao et al. (2005) consider a decentralized dealer network in which each dealer can share its inventory with the others. Each dealer gives the high

priority to its own customers and the low priority to other dealers. They assume the critical level rationing and use the threshold clearing mechanism that Dehpande et al. (2003) propose. They analyze the system by constructing an inventory sharing game.

We conclude the chapter with Table 2.1. It summarizes the stock rationing literature. The literature is classified on the basis of the backordering and lost sales cases. Moreover, the works on the production environment, i.e. capacitated replenishment channel, are also provided.

Table 2.1 Stock Rationing Literature

	Periodic review	Continuous review	Production environment
Backordering	Veinott(1965) Topkis (1968) Kaplan (1969) Nahmias and Demmy (1981) Moon and Kang (1998) Atkins and Katircioglu (1995)	Nahmias and Demmy (1981) Teunter and Haneveld (1996) Dekker et al. (1998) Moon and Kang (1998) Deshpande et al. (2003) Melchiors (2003) Kocaga (2004) Arslan et al. (2005) Zhao et al. (2005)	Ha (1997b) Ha (2000) Vericourt (2002) Gayon et al. (2005)
Lost sales	Topkis (1968) Evans (1968) Cohen et .al (1989)	Melchiors et al. (2000) Dekker et al. (2002)	Ha (1997a)

Chapter 3

NOTES ON THE CRITICAL LEVEL POLICY AND RELATED LITERATURE

In this chapter, we present our observations on the critical level policy with two demand classes and backordering under continuous review. These observations can also be extended easily to multiple demand class systems. We clarify many ambiguities resulting from the literature and position the contributions of Dekker et al (1998) and Deshpande et al. (2003), the two important works in the area.

For the stock rationing problem, even with the exponential lead times, i.e. the simplest setting, there is no work in the literature that characterizes the optimal policy structure. Related literature concentrates on the critical level policy, which assumes static threshold level. Although it is known that critical level policy is not optimal, having a fixed threshold level makes it the easiest policy to analyze and implement. Moreover, in literature there is no agreement on the backorder clearing mechanism to be used within the critical level policy. An important reason for assuming clearing mechanisms other than the priority clearing is to make the analytical analysis possible. However, there is no work that clarifies the connection between the critical level policy and the priority clearing mechanism. Although Deshpande et al.

(2003) state that they propose a different clearing mechanism because they could not get any analytical results with the priority clearing mechanism, they do not assess the necessity of priority clearing when the critical level policy is used.

We organize this chapter in two sections. In section 3.1, we explain why the priority clearing mechanism should be the natural consequence of the critical level policy. In section 3.2, we discuss different clearing mechanisms considered in the literature and also show that when $Q=1$, the fill rate expressions of Deshpande et al. (2003) turn out to be the expressions given in Dekker et al (1998).

Before proceeding with the sections of the chapter, we provide the following notation:

β_i = $P\{\text{an arriving class i customer is served immediately}\}$, i.e. fill rate for class i and i = 1,2.

β_i^{PC} = fill rate for class i when the critical level policy is used with the priority clearing mechanism, i = 1,2.

β_i^{AC} = fill rate for class i when the critical level policy is used with an alternative clearing mechanism, i = 1,2.

γ_i = average backorder time per customer for class i, i = 1,2.

γ_i^{PC} = average backorder time per customer for class i when the critical level policy is used with the priority clearing mechanism, i = 1,2.

γ_i^{AC} = average backorder time per customer for class *i* when the critical level policy is used with an alternative clearing mechanism, *i* = 1,2.

$E[TC] = A\frac{\lambda_1+\lambda_2}{Q} + h\bar{I} + \pi_1(1-\beta_1)\lambda_1 + \pi_2(1-\beta_2)\lambda_2 + \hat{\pi}_1\gamma_1\lambda_1 + \hat{\pi}_2\gamma_2\lambda_2$ is the expected total cost rate where A is the fixed ordering cost, h is the unit holding cost and $\bar{I}$ is the average inventory.

In this chapter and also in chapter 5, we use simulation results for comparison and performance evaluation purposes. The simulation runs are controlled by the total number of arrivals. We run the simulation until 500,000 total arrivals in order to observe the steady state behaviors of the policies. To verify the accuracy of our simulation, we simulate the critical level policy with the following parameter set; $r = 5, Q = 1, K = 3, \lambda_1 = 3, \lambda_2 = 2, L = 1$ where L is the leadtime. With the significance level 0.05 and 10 replications, for the fill rate for class 1 we obtained a confidence interval (0.92024, 0.920455) and for the fill rate for class 2 we obtained (0.124471, 0.124742). These confidence intervals are small enough to allow us to use the average of 10 replications for each case.

3.1 The Priority Clearing Mechanism vs. Other Clearing Mechanisms

It is possible to define many different clearing mechanisms under the critical level policy. Each policy results in different service levels and expected cost rate due to allocating the order quantity in different ways between increasing the stock level and clearing the backorders of different customer classes. However, the critical level policy provides service to class 2 when the on hand stock is above K and reserves all the stock below K for class 1 customers. Thus, we think that the natural clearing mechanism of the critical

level policy should be the one that postpones the clearance of class 2 backorders until the on hand inventory reaches K, i.e. until filling the reserve stock of class 1 customers. This clearing mechanism exactly corresponds to the priority clearing. Hence, to make it a well-defined policy, the critical level policy should be defined with the priority clearing mechanism. With any other clearing mechanism, K does not really corresponds to a threshold level under which all the stock reserved for class 1.

We verify the idea mentioned above by considering the effects of clearing mechanisms on the service levels, which are the performance measures of the system, for any fixed set of input policy parameters. It is obvious that upon arrival of replenishment order, the system should first clear class 1, high priority, backorders if there is any, because the time dependent backorder cost of class 1 is higher, that is $\hat{\pi}_1 > \hat{\pi}_2$. After the clearance of class 1 backorders, if the order quantity is not depleted, one can choose to clear some class 2 backorders before inventory level reaches K or s/he can first choose to increase the stock level up to K and then clear class 2 backorders. The latter one corresponds to the priority clearing mechanism. If any class 2 backorder is cleared before increasing the stock level up to K, let us call it alternative clearing, the resulting fill rate for class 1, β_1^{AC}, is less than β_1^{PC}. This is so because when the priority is given to clearing some or all class 2 backorders, the remaining order quantity may not be enough to increase the inventory up to K. Therefore, compared to the priority clearing, the number of future class 1 demands that finds the system in stockout increases.

In addition to the decrease in the fill rate for class 1, clearing some class 2

backorders before the inventory hits K does not provide any increase in the fill rate for class 2, i.e. $\beta_2^{AC} = \beta_2^{PC}$. This is so because fill rate gives the ratio of customers that are served immediately when they arrive. Under the critical level policy, class 2 arrivals are satisfied when the on hand stock level is between K and $(r+Q)$. Therefore, not only the priority clearing mechanism, any mechanism that requires the clearing of all backorders before increasing the stock level above K results in a class 2 fill rate same as β_2^{PC}, i.e. the fill rate for class 2 is independent from the clearing mechanism if the stock level increases above K after the clearance of all backorders. These kinds of policies fully allocate the on hand inventory above K to the future arrivals of both classes.

Deviating from the priority clearing mechanism and using an alternative clearing provides some decrease in the total backorder time of class 2 demands. Giving priority to clear some class 2 backorders decreases average backorder time for class 2, i.e. $\gamma_2^{AC} < \gamma_2^{PC}$. However, as we discussed above, compared to the priority clearing, such a clearing mechanism decreases the fill rate for class 1. Thus, there are more backorders from class 1 and so average backorder time for class 1 increases, i.e. $\gamma_1^{AC} > \gamma_1^{PC}$. Increasing average backorder time for class 1 is not rational because $\hat{\pi}_1 > \hat{\pi}_2$. Therefore, from the perspective of service levels, the fill rates and the average backorder times, priority clearing should be the natural consequence of critical level policy.

3.2 Notes on Dekker et al. (1998) and Deshpande et al. (2003)

Nahmias and Demmy (1981) work on the critical level policy for continuous review systems with backordering for the first time in the literature. They assume Poisson arrivals of two demand classes, deterministic leadtime and at most one outstanding replenishment order. The two most important works that generalize the setting of Nahmias and Demmy (1981) are Dekker et al (1998) and Deshpande et al. (2003). Without having any restriction on the number of outstanding orders, Dekker et al (1998) analyze the $(K, S-1, S)$ policy and Deshpande et al. (2003) consider the more general (K, r, Q) policy. However, there are some ambiguities in these works. To point out those ambiguities and to set the connection between these two works we present a deeper analysis in this section.

For the $(K, S-1, S)$ policy, i.e. the order quantity Q is 1, Dekker et al. (1998) discuss three clearing mechanisms including the priority clearing after deriving expressions for the fill rates β_1 and β_2. Similar to the priority clearing, the other two mechanisms that Dekker et al. (1998) suggest also use the arriving order quantity to clear the oldest class 1 backorder first. However, if there is no class 1 backorder and the stock level is below *K*, one of the mechanisms gives the priority to clearing the oldest class 2 backorder instead of increasing the stock level. But the other mechanism gives the priority to increasing the stock if a class 1 demand triggered the arriving order; otherwise it gives the priority to clear the oldest class 2 backorder. If the stock level is at K, then both of the mechanisms clear the oldest class 2 backorder if there is any, if not, they increase the stock. Thus, as in the case of the priority clearing, the stock level cannot be increased above *K* before clearing all the backorders. Therefore, all the three mechanisms result in the

same fill rate for class 2 as we discussed in Section 3.1.

The fill rate expression for class 2 that Dekker et al. (1998) provide is the following:

$$\beta_2 = \sum_{x=0}^{S-K-1} p(x, \lambda L) \tag{3.1}$$

In (3.1) $p(x, \lambda L)$ is the Poisson probability of x arrivals in the leadtime, and $\lambda = \lambda_1 + \lambda_2$

The logic behind equation (3.1) is as follows: We know that the inventory position is at the order-upto-level S for at any time point t. As in the analysis of Hadley and Whitin (1963), to satisfy a class 2 demand that arrives at $t + L$, the leadtime demand should be at most inventory position at t minus $K - 1$.

Dekker et al. (1998) get the fill rate expressions without considering any clearing mechanism. They claim that the expression given in equation (3.1) is exact because the fill rate for class 2 is independent of the clearing mechanism. However, we should point out that β_2 also depends on the clearing mechanism. The independence of β_2 only holds for the clearing mechanisms that clear all class 2 backorders before increasing the inventory level above *K*. The logic behind the equation (3.1) is only valid for the clearing mechanisms that belong to this category. As noted before, the clearing mechanisms of Dekker et al. (1998) all belong to this category. However, it is easy to show that their claim is not true in general by considering a counter example. If we define a mechanism in such a way that we clear the oldest class 2 backorder after the on hand inventory level reaches *K+1*, then under this clearing the fill rate for class 2 would be

certainly greater than the expression given in equation (3.1). Because the ratio of the time that the inventory level is above K increases due to the postponement of clearing class 2 backorders. Then, the probability of satisfying an arriving class 2 demand, i.e. β_2, increases.

The expression that Dekker et al. (1998) suggest for the class 1 fill rate is

$$\beta_1 = \beta_2 + \sum_{i=0}^{K-1} \int_0^L \lambda^{S-K} \frac{y^{S-K-1}}{(S-K-1)!} e^{-\lambda y} \frac{e^{-\lambda_1 (L-y)} [\lambda_1 (L-y)]^i}{i!} dy \qquad (3.2)$$

The logic behind equation (2) is as follows: Class 1 demands are satisfied in the region that class 2 demands are satisfied. In addition, a class 1 demand that arrived a leadtime later than the time the system observed will be filled if there will be at least one stock on hand, i.e. there should be at most $K-1$ class 1 demands within the leadtime. Equation (3.2) tries to capture this fact by conditioning on the "hitting time" of the critical level. Dekker et al. (1998) claim that "hitting time" is $(S-K)$ stage Erlang random variable with parameter $\lambda_1 + \lambda_2$.

As Dekker et al. (1998) state that the expression in (3.2) is independent of the clearing mechanism and so it is an approximation of the realized fill rate for class 1. However, there are some other problems related to this approximation. It is true that at any time point the inventory position is S, but rationing decision depends on the on-hand stock level. Therefore, their "hitting time" is not the real hitting time of critical level K if the on hand level is not S at the time when the system is observed. The inventory level hits K after $(S-K)$ demand arrivals only if it starts at S. Moreover, even though it is not mentioned, the expression in (3.2) assumes that rationing continues until the end of the leadtime once it starts. This means that the

effect of incoming replenishment orders within the leadtime are ignored. Replenishment orders may increase the inventory level above K. Therefore, within the leadtime system may again start to satisfy class 2 demands. No replenishment order arrival within the leadtime is only possible if the inventory level is at S when the system is observed, i.e. no outstanding orders. This is equivalent to at-most-one-order-outstanding assumption although the authors claim the otherwise. Therefore, if the steady state probability of being at level S decreases, the approximation gets worse dramatically. By increasing the traffic rate this situation can be observed.

Before the comparison of the realized fill rate for class 1 and the approximation of Dekker et al. (1998), let us analyze the clearing mechanism and the fill rate expressions of Deshpande et al. (2003). They consider (K,r,Q) policy and proposes the threshold clearing mechanism that allow clearing some class 2 backorders before the inventory level reaches to K. Later Deshpande and Ryan (2005) also use the threshold clearing mechanism. Under this mechanism Deshpande et al. (2003) define a clearing position that starts at $r+Q$ when an order is placed. Up to the threshold level K clearing position decreases with the total demand rate $\lambda = \lambda_1 + \lambda_2$. Then, it continuous decreasing at rate λ_1. When a replenishment order of size Q arrives, the rules to apply the threshold clearing are as follows:

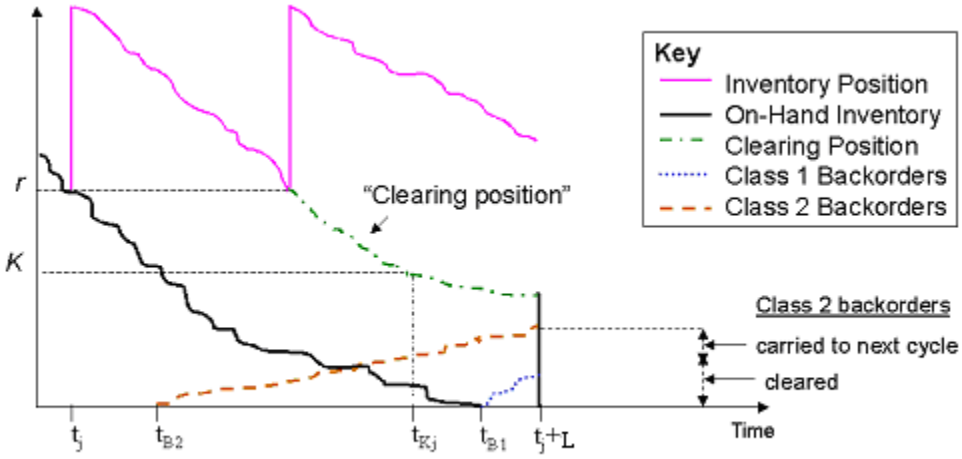
1. Clear all class 1 and class 2 backordered demand that arrived before the clearing position hits K on the basis of FCFS rule.

2. Clear any remaining class 1 backordered demand if possible with the remaining order quantity. Continue to backorder all class 2 demand that arrive after the clearing position hits K.

Figure 3.1 summarizes the threshold clearing mechanism for a typical cycle. At time t_j, j^{th} order is placed. At t_{B2} on hand stock hits K and the system start to backorder the class 2 demands. At t_{B1} the on hand stock is depleted. And t_{Kj} is the time that the clearing position of this specific order hits K. After the completion of the clearing procedure, clearing position and the on hand inventory meet at the same level.

With the threshold clearing mechanism, Deshpande et al. (2003) enables to get the exact steady state characteristics of the system. However, in contrast to the priority clearing, threshold clearing fills some class 2 backlogged demand before filling all class 1 backlogged demands.

FIGURE 3.1 Threshold Clearing Mechanism



As in the case of Figure 3.1, if an order is placed at time t, it will arrive at time $t+L$. Deshpande et al. (2003) define $D(t,t+L)$ to denote the total demand between the placement and the arrival of the replenishment order. Then, if $D(t,t+L) \le r+Q-K$ then all type of backorders are cleared and on hand stock is increased to $r+Q-D(t,t+L)$, which is greater or equal to

K. This is so because, $r - D(t, t+L)$ corresponds to the level after subtracting all satisfied and backordered demands, and this level plus the order quantity Q carried the stock level to at least K. This implies that if the stock level is at or above K, there is no backorders of any type. As in the case of the clearing mechanisms that Dekker et al. (1998) suggested, from the discussion in Section 3.1, the threshold clearing mechanism must result in the same fill rate for class 2 with the priority clearing mechanism. Consequently, the expression in Equation (3.3), which Deshpande et al. (2003) derive under the threshold clearing, must be exactly the same for the priority clearing mechanism.

$$\beta_2 = \frac{1}{Q} \sum_{y=r+1}^{r+Q} \sum_{x=0}^{y-K-1} p(x, \lambda L) \tag{3.3}$$

Therefore, we can get (3.3) directly without assuming any specific clearing mechanism: As in the analysis of Hadley and Whitin (1963), by conditioning on the inventory position at any time point t, which is uniform on $[r+1, r+Q]$, we can get the distribution of on hand stock level by considering the lead time demand. Furthermore, to satisfy a class 2 demand that arrives at $t+L$, the leadtime demand should be at most inventory position at *t* minus *K-1*. This logic is only valid for the clearing mechanisms that our observation is applicable, i.e. β_2 is independent from the clearing mechanism if the stock level increases above K after the clearance of all backorders.

Note that, by assuming $Q = 1$ and $r = S - 1$, we can get equation (3.1), β_2 of Dekker et al. (1998), from equation (3.3). This is a verification of our observation. Moreover, we test the validity of our observation using simulation results. For 5 cases, which illustrate totally different scenarios by

considering different arrival rates, leadtimes and (K, r, Q) values, Table 3.1 shows the simulation results of β_2 under the priority clearing and the expression given in equation (3.3).

TABLE 3.1 Comparison of the class 2 fill rates that are obtained with threshold clearing and priority clearing

$(\lambda_1, \lambda_2, L)$	(K, r, Q)	β_2 (threshold clearing)	β_2 (priority clearing)
(0.3, 0.7, 8)	(1,10,20)	0.964	0.964
(0.3, 0.7, 8)	(5,15,20)	0.978	0.978
(1, 1, 1)	(1,1,1)	0.135	0.135
(3, 2, 1)	(5, 3, 4)	0.011	0.011
(4, 6, 1)	(2, 8, 5)	0.344	0.344

As the last observation, it is important to note that when $Q = 1$, the fill rate expression for class 1 that Deshpande et al. (2003) provides for the threshold clearing turns out to be the approximation of Dekker et al. (1998), which is given in equation (3.2). Interestingly, Dekker et al. (1998) construct the expression in (3.2) without assuming any clearing mechanism, but it gives the exact β_1 under threshold clearing mechanism when $Q = 1$.

When $Q = 1$ and $r = S - 1$, β_1 expression of Deshpande et al. (2003) is

$$\beta_1 = 1 - \sum_{x=S}^{\infty} \sum_{j=0}^{x-S} b(\alpha_1; x - S + K; K + j)\, p(x, \lambda L) \tag{3.4}$$

In (3.4), $b(\alpha_1; x - S + K; K + j) = \dfrac{(x - S + K)!}{(K + j)!(x - S - J)!} \alpha_1^{K+j} (1 - \alpha_1)^{x-S-j}$, i.e. a binomial probability, and $\alpha_1 = \dfrac{\lambda_1}{\lambda}$.

The expression in equation (3.4) means that β_1 is equal to one minus the

probability that the leadtime demand for both classes is at least S and at least K of the leadtime demand is from class 1. But, we can interpret equation (3.4) in a different way. β_1 of (3.4) is composed of two parts. First part is the probability that the leadtime demand for both classes is at most $S-K-1$. The second part is the probability of $S-K$ total demand within the leadtime, i.e. probability of hitting K within the leadtime, plus at most *K-1* class 1 demands in the remaining part of the leadtime. Then at least one unit of inventory is available at the end of the leadtime, i.e. the total demand that decreases the inventory is $S-1=(S-K)+(K-1)$. This interpretation of (3.4) is exactly what the equation (3.2) says, which is the class 1 fill rate approximation of Dekker et al. (1998).

As mentioned before, Dekker et al. (1998) derive the approximate β_1 expression, equation (3.2), without assuming any clearing mechanism and Deshpande et al. (2003) work with the threshold clearing mechanism because they are "unable to perform an analysis under the priority clearing". Moreover, for some parameter sets they test the performance of the threshold clearing mechanism by comparing it with the simulation results of priority clearing. Therefore, for $Q=1$, we compare their class 1 fill rate expressions, which are equal to each other, with simulation results of the system under the critical level policy with priority clearing mechanism. Table 3.2 demonstrates this comparison for different arrival rates. We choose $L=1$ and $K=1$, $r=4$. As we noted before, if the steady state probability of being at level S decreases, the approximation of Dekker et al. (1998) gets worse dramatically. And by increasing the traffic rate this situation can be observed. Table 3.2 illustrates this fact. Moreover it shows that equation (3.2) and (3.4) gives the same results as expected.

CHAPTER 3 NOTES ON THE CRITICAL LEVEL POLICY

As seen from Table 3.2, for the priority clearing mechanism, equations (3.2) and (3.4) always underestimate the class 1 fill rate. We already stated that the threshold clearing mechanism results in a lower fill rate for class 1 compared to the priority clearing mechanism. However, we can also explain the situation by just considering the approximate fill rate expression for class 1 that is provided by Dekker et al. (1998). Equation (3.2) assumes that the inventory level is at S when the system is observed and it decreases with rate $\lambda_1 + \lambda_2$ until K, and then decreases with λ_1. Equivalently, this is same as starting at any inventory level and assuming that the replenishment orders clear all the backorders within the leadtime and increases the inventory level up to S. This is a direct application of the logic that is used to derive the steady-state distribution of the inventory level for the $(S-1, S)$ policy without rationing, i.e. inventory level is S minus the leadtime demand, because if no demand arrives replenishment orders carries the inventory level to S. However, for the $(K, S-1, S)$ policy, replenishment orders do not clear class 2 backorders and increase the inventory level if it is below K. Increasing the stock instead of clearing a class 2 backorder increases the fill rate for class 1. Therefore, by assuming that all the backorders are cleared within the leadtime, Equation (3.2) underestimates the fill rate for class 1.

TABLE 3.2 Comparison of the class 1 fill rates

(λ_1, λ_2)	β_1 (Dekker et al. (1998))	β_1 (Deshpande et al. (2003))	β_1 (Simulation)
(1,1)	0.968	0.968	0.973
(1.5,1.5)	0.881	0.881	0.907
(2,1.5)	0.799	0.799	0.843
(8,1.5)	0.050	0.050	0.211

CHAPTER 3 NOTES ON THE CRITICAL LEVEL POLICY

Before concluding the chapter we summarize our observations:

1. Priority clearing should be the natural clearing mechanism of the critical level policy. Any alternative mechanism negatively affects the service levels for class 1 without any increase in the fill rate for class 2.

2. All clearing mechanisms that clear all backorders before increasing the inventory level above K result in the same fill rate value for class 2.

3. Class 2 fill rate expression of Dekker et al. (1998) is exact only for the clearing mechanisms that clear all backorders before increasing the inventory level above K.

4. The approximation of Dekker et al. (1998) for the fill rate for class 1 is valid if we assume at-most-one-order-outstanding.

5. For $Q = 1$, fill rate expressions of Deshpande et al. (2003) turn out to be the expressions that Dekker et al. (1998) provide.

6. For the priority clearing mechanism, class 1 fill rate expressions provided by Deshpande et al. (2003) and Dekker et al. (1998) always underestimate the realized fill rate for class 1.

Chapter 4

AN EMBEDDED MARKOV CHAIN APPROACH

In this chapter, we present a new method for the exact analysis of continuous-review lot-per-lot inventory systems with backordering under the critical level rationing policy on two priority classes. This method is based on the observation that the state of the inventory system sampled at multiples of the supply leadtime evolves according to a Markov chain. Our analysis yields the steady-state distribution for the inventory system, which can be used to obtain any long-run performance measure. An exact steady-state analysis for the inventory system was not available up to this point.

The one-step transition probabilities of the embedded Markov chain corresponding to the inventory system are generated using a recursive procedure we develop in Section 4.1. This procedure is based on four recursion equations that are valid in different regions of the state space of the Markov chain and two equations for the boundaries of the state space. Section 4.2 is devoted to steady-state analysis of the chain. In this section, we show that we can get steady-state probabilities of interest with desired accuracy by considering transition probabilities corresponding to a subset of the state space. The application of this technique is mandatory for our problem since the state space for the embedded Markov chain is infinite. Finally, we demonstrate that the technique converges to acceptable accuracy levels fairly quickly by reporting the results of the technique on an instance

of the inventory system considered.

4.1. The Embedded Markov Chain

We assume inventory for an item is held and replenished over time in order to keep up with the demand from two customer classes, which occur according to two Poisson processes with rates λ_1 and λ_2, accordingly. This means that the total demand also follows a Poisson process with rate $\lambda = \lambda_1 + \lambda_2$. Any unmet demand is backlogged. The inventory policy is a lot-per-lot $(S-1, S)$ with the critical level rationing and the priority clearing mechanism.

In the analysis of continuous-review inventory systems, the state of the system is usually selected as the inventory level. But under rationing policy, there may be class 2 backorders when the inventory level is under the support level. Thus, one needs a two-dimensional state-space to keep track of the inventory level and the number of class 2 backorders.

To define the state of the system not in terms of the inventory level, but in terms of the number of outstanding replenishment orders lends itself better to analysis. Therefore, we define the state of the system at time t as $(X(t), B(t))$, where $X(t)$ is the number of outstanding replenishment orders at time t and $B(t)$ is the number of class 2 backorders at time t. Under a lot-for-lot inventory policy, $X(t)$ is also equal to the number of demand arrivals of both types in $(t-L, t]$, since each demand arrival triggers a replenishment order. Notice that the inventory level at time t,

$$I(t) = S - X(t) + B(t), \tag{4.1}$$

since demand during $(t-L,t]$ reduces the inventory level from the inventory position at time $t-L$, which is *S*, given that it does not correspond to a class 2 backorder. There is no backorder if the inventory level is above the support level *K*, since all backorders need to be cleared at the support level, i.e.

$$B(t) = 0 \ \ if \ X(t) < S - K \tag{4.2}$$

If the inventory level is at *K* or lower, only class 2 demands that arrive after the inventory level hits *K* would be backordered. Even if all the demand arrivals in $(t-L,t]$ belong to class 2, the maximum number of backorders would be $X(t)-(S-K)$. Thus,

$$0 \le B(t) \le X(t)-(S-K) \quad if \ X(t) \ge S-K \tag{4.3}$$

Conditions (4.2) and (4.3) specify the feasible states and thereby the state space of the embedded Markov chain, while Equation (4.1) specifies the inventory level corresponding to each state of the state space.

Given that we know the state of the system at time *t*, it is possible to derive the probability that the system reaches a certain state at time *t*+*L*. If we derive this probability for all feasible states at time *t*, and at time *t*+*L*, then we obtain an embedded Markov chain for the inventory system. These probabilities are the one-step transition probabilities of the Markov chain. They determine the probabilistic evolution of the inventory system at multiples of leadtime. One should note that the original continuous-time process describing the evolution of the states at any point in time is regenerative. The process regenerates itself every time there is no outstanding replenishment order in the inventory system, i.e., when $X(t) = 0$ and $B(t) = 0$. Since the underlying continuous-time process is regenerative,

the process is ergodic and the limiting distribution of the process exists (See Stidham 1974). When the number of transitions for the embedded Markov chain tends to infinity, the probabilities observed will be the probabilities for the continuous-time process as $t \to \infty$. Thereby, the limiting distribution of the embedded Markov chain has to be the same with the underlying continuous-time process, i.e.,

$$\lim_{t\to\infty} P\{X(t)=x, B(t)=b\} = \lim_{n\to\infty} P\{X(nL)=x, B(nL)=b\}.$$

Thus, the limiting distribution of the embedded Markov chain is sufficient for statistical characterization of the inventory system in the long run.

One can observe that one of our state variables *X(t)*, sampled at multiples of leadtime, evolves itself according to an embedded Markov chain. Moreover, the one-step transition probabilities of this embedded Markov chain are independent of the origin state, i.e.

$$P\{X(t+L)=x_L \mid X(t)=x_0\} = P\{D_{(t,t+L]}=x_L\} = e^{-\lambda L}\frac{(\lambda L)^{x_L}}{x_L!} \quad for\ x_L = 0,1,2,... \tag{4.4}$$

The evolution of *X(t)*, number of outstanding replenishment orders, is fully independent of the rationing policy. The result stated in (4.4) is the basis of the steady-state analysis of $(S-1,S)$ inventory systems (see Hadley and Whitin (1963) pages 204-205). A direct implication of (4.4) is that the distribution of *X(t)* converges to its limiting distribution at $t = L$. Thus, we are able to decouple one of the dimensions of the two-dimensional chain, and solve it independently. This simplifies our analysis considerably.

We can express all first step probabilities as

$$P\{X(t+L)=x_L, B(t+L)=b_L \mid X(t)=x_0, B(t)=b_0\}$$

The probabilities that relate to reaching an inventory level above the support level K can be obtained directly from (4.4) as

$$P\{X(t+L)=x_L, B(t+L)=0 \mid X(t)=x_0, B(t)=b_0\} = P\{D_{(t,t+L]}=x_L\} = e^{-\lambda L}\frac{(\lambda L)^{x_L}}{x_L!} \quad (4.5)$$
$$\text{for } 0 \le x_L \le S-K, \text{ for all feasible } (x_0, b_0) \text{ pairs,}$$

since there will be no backorders at time $t+L$.

One needs considerably more effort in order to obtain other one-step transition probabilities. Since we know the distribution of $X(t+L)$ and the fact that the distribution is independent of $X(t)$ and $B(t)$, we can use this for our objective in

$$P\{X(t+L)=x_L, B(t+L)=b_L \mid X(t)=x_0, B(t)=b_0\} =$$
$$\sum_{x_L=0}^{\infty} P\{B(t+L)=b_L \mid X(t+L)=x_L, X(t)=x_0, B(t)=b_0\} e^{-\lambda L}\frac{(\lambda L)^{x_L}}{x_L!}. \quad (4.6)$$

Thus, we need to compute

$$P\{B(t+L)=b_L \mid X(t+L)=x_L, X(t)=x_0, B(t)=b_0\} \quad (4.7)$$

for all feasible state pairs (x_0, b_0) and (x_L, b_L). This is the probability that there are b_L units of backorder at time $t+L$, given that there are b_0 units of

backorder at time t, x_0 demand arrival occurs in $(t-L,t]$, and x_L demand arrivals occurs in $(t,t+L]$. Since arrivals occur according to a Poisson process, the unordered arrival times in $(t-L,t]$ are x_0 independent random variables with uniform distribution on $(t-L,t]$ and the unordered arrival times in $(t,t+L]$ are x_L independent random variables with uniform distribution on $(t,t+L]$. Each demand arrival in $(t-L,t]$ triggers a replenishment order that arrives exactly in L units of time. This means that replenishment order arrival times in $(t,t+L]$ are x_0 independent random variables with uniform distribution on $(t,t+L]$. Moreover, since demand arrival times in $(t-L,t]$ and $(t,t+L]$ are independent, the x_L demand and the x_0 replenishment arrival times in $(t,t+L]$ are all independent from each other.

Since the number of order and replenishment arrivals during $(t,t+L]$ is known, it is the order of the replenishment and demand arrivals, and the class of the demand arrivals, which determines the number of backorders reached at time $t+L$. Unfortunately, it is not possible to obtain a closed form expression for the probability expression (4.7). Yet, it is still possible to compute these probabilities using a recursive procedure. This procedure is based on a related probability expression, which can be expressed as

$$P\left\{B\left(t+L\right)=b_L \mid X\left(t+L\right)=x_L, B\left(t'\right)=b, Y\left(t'\right)=y, Z\left(t'\right)=z\right\} \quad \text{for } t\le t'\le t+L \tag{4.8}$$

where *Y(t')* is the number of demand arrivals in $(t',t+L]$, and *Z(t')* is the number of replenishment arrivals in $(t',t+L]$, whose occurrence times are

all independent and identically distributed on $(t', t+L]$. The reader should note that the probabilities in (4.8) do not depend on *t'*. This independence is due to fact that once the number of arrivals during $(t', t+L]$ is known, the duration of the period does not change anything. We can now express (4.7) in terms of (4.8) as

$$P\{B(t+L)=b_L \mid X(t+L)=x_L, X(t)=x_0, B(t)=b_0\} = P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b_0, Y(t')=x_L, Z(t')=x_0\} \quad \text{for } t \le t' \le t+L. \tag{4.9}$$

The conditions of the conditional probability (4.8), do not include the number of outstanding replenishment orders at time *t'*. But, given $X(t+L)=x_L, Y(t')=y, Z(t')=z$, this quantity is readily determined by

$$X(t') = X(t+L) + Z(t') - Y(t') = x_L + z - y \tag{4.10}$$

The logic behind Equation (4.10) can be explained as follows: In order to find the number of outstanding replenishment orders at the end of the period $(t', t+L]$, i.e., $X(t+L)$, one should add the difference between the number of demand arrivals in $(t', t+L]$ (which trigger new replenishment orders) and the number of replenishment arrivals in $(t', t+L]$ (which clear the outstanding replenishment orders) to the number of outstanding replenishment orders at the beginning of the period.

The recursive procedure devised to compute the probabilities defined by (4.8) is based on the fact that the probabilities conditioned on the number of arrivals that occur in $(t', t+L]$, can be written in terms of the same kind of probabilities conditioned on fewer arrivals during the same period. In order to write these relations, we need to consider what happens next depending on

the nature of the first arrival in $(t', t+L]$. There are three possible events that can take place: a replenishment order arrival, a class 1 demand arrival, and a class 2 demand arrival. Since the arrivals in $(t', t+L]$ will be uniformly distributed on $(t', t+L]$, the probability that a replenishment order arrives first is the proportion of outstanding replenishments to the total number arrivals, i.e. $Z(t')/\big(Z(t')+Y(t')\big)$. The complement of this probability, $Y(t')/\big(Z(t')+Y(t')\big)$, is the probability that a demand arrival occurs first. This demand arrival will be of class 1 with probability $p_1 = \lambda_1/(\lambda_1+\lambda_2)$ and will be of class 2 with probability $p_2 = 1-p_1$.

The effect of this first arrival depends on the number of outstanding replenishment orders at that time as a result of rationing policy. If the number of outstanding replenishment orders is less than *S*-*K*, then there is no backorder in the system and a demand arrival does not cause any backorder either, i.e.,

if $0 \le x_L + z - y \le S-K-1$, then

$$P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=0, Y(t')=y, Z(t')=z\} = \tag{4.11}$$
$$\frac{z}{z+y} P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=0, Y(t')=y, Z(t')=z-1\}$$
$$+\frac{y}{z+y} P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=0, Y(t')=y-1, Z(t')=z\}.$$

If the number of outstanding replenishment orders is exactly *S*-*K*, again there is no backorder at the system. If a replenishment order arrives first, then the state will move to the region of Equation (4.11). If a class 1 demand

arrives first, the inventory level drops by one; and if the class 2 demand arrives first, that demand is backordered and the inventory level remains the same, i.e.,

if $x_L + z - y = S - K$, then

$$P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = 0, Y(t') = y, Z(t') = z\} =$$
$$\frac{z}{z+y} P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = 0, Y(t') = y, Z(t') = z-1\} \quad (4.12)$$
$$+ \frac{y}{z+y} p_1 P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = 0, Y(t') = y-1, Z(t') = z\}$$
$$+ \frac{y}{z+y} p_2 P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = 1, Y(t') = y-1, Z(t') = z\}.$$

If the number of outstanding replenishment orders is greater than *S*-*K* and the inventory level is still at *K*, then the system is at the clearing position. This means, if a replenishment order arrives first, one unit of class 2 backorder is cleared. If a class 1 demand arrives first, the inventory level drops by one; and if a class 2 demand arrives first, that demand is backordered and the inventory level remains the same, i.e.,

if $x_L + z - y \geq S - K + 1$ and $x_L + z - y - b = S - K$, then

$$P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b, Y(t') = y, Z(t') = z\} =$$
$$\frac{z}{z+y} P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b-1, Y(t') = y, Z(t') = z-1\} \quad (4.13)$$
$$+ \frac{y}{z+y} p_1 P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b, Y(t') = y-1, Z(t') = z\}$$
$$+ \frac{y}{z+y} p_2 P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b+1, Y(t') = y-1, Z(t') = z\}.$$

Finally, if the number of outstanding replenishment orders is greater than S-K and the inventory level is less than K, than the replenishments do not cause any clearing, but increase the inventory level. The demand arrivals either decrease the inventory level or are backordered depending on their class, i.e.,

if $x_L + z - y \geq S - K + 1$ and $x_L + z - y - b \geq S - K$, then

$$
\begin{aligned}
&P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b, Y(t') = y, Z(t') = z\} = \\
&\frac{z}{z+y} P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b, Y(t') = y, Z(t') = z-1\} \qquad (4.14) \\
&+ \frac{y}{z+y} p_1 P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b, Y(t') = y-1, Z(t') = z\} \\
&+ \frac{y}{z+y} p_2 P\{B(t+L) = b_L \mid X(t+L) = x_L, B(t') = b+1, Y(t') = y-1, Z(t') = z\}.
\end{aligned}
$$

So far, we have defined four different regions in our state space and we have derived four equations ((4.11)-(4.14)), one for each region, which expresses the probabilities in (4.8) as a function of probabilities of the same type, which correspond to smaller number of arrivals. This means if we know the probabilities corresponding to fewer arrivals, then we can compute those probabilities corresponding to more arrivals. This fact constitutes the basis of the recursive procedure we propose. All that is needed to complete the procedure are the boundary equations, which determine the probabilities corresponding to no replenishment arrival and to no demand arrival. From these probabilities, all other probabilities are computed in a recursive fashion.

The first boundary condition determines the probabilities corresponding to no demand arrival. Under this condition, only a certain number of replenishments arrive in $(t', t+L]$. These replenishments first increase the

inventory level to the support level. If there are more replenishment orders, they clear the class 2 backorders. If there are still more replenishments they continue increasing the inventory level, i.e.,

$$P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b, Y(t')=0, Z(t')=z\}$$
$$=\begin{cases} 1\{b_L=0\}, & \text{if } x_L \le S-K \\ 1\{b_L=b\}, & \text{if } x_L > S-K \text{ and } x_L-b \ge S-K \\ 1\{b_L=b-((S-K)-(x_L-b))\}, & \text{if } x_L > S-K \text{ and } x_L-b < S-K. \end{cases} \quad (4.15)$$

$1\{\cdot\}$ is the indicator function. It is 1 if the condition holds and 0 otherwise.

The second boundary condition determines the probabilities corresponding to no replenishment arrival. Under this condition, only a certain number of demands arrive in $(t', t+L]$. These demands first decrease the inventory level to the support level. If there are more demands to arrive, class 1 demands decrease the inventory and class 2 demands increase the number of class 2 backorders. The second boundary condition does not yield a close form expression unlike the first. We again obtain an equation that needs to be solved recursively, which is

$$P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b, Y(t')=y, Z(t')=0\} =$$
$$\begin{cases} P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b, Y(t')=y-1, Z(t')=0\}, \\ \qquad\qquad \text{if } x_L-y \le S-K-1 \\ p_1 P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b, Y(t')=y-1, Z(t')=0\} \\ +p_2 P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b+1, Y(t')=y-1, Z(t')=0\}, \\ \qquad\qquad \text{if } x_L-y \ge S-K. \end{cases} \quad (4.16)$$

The probability $P\{B(t+L)=b_L \mid X(t+L)=x_L, B(t')=b, Y(t')=0, Z(t')=0\}$

is already determined by Equation (4.15). Starting with this probability, one can obtain the probabilities on the second boundary by increasing $Y(t')$ one by one.

With Equations (4.11)-(4.16), one can start with the one-step transition probabilities corresponding to $Y(t')=0$, and $Z(t')=0$; and then obtain the rest of the one-step transition probabilities increasing $Y(t')$ and then $Z(t')$ one by one in a recursive fashion. Thus, all the transition probabilities can be generated with this procedure. The only problem is that the state space of the embedded Markov chain is infinite and it would take infinite amount of time to generate all the elements of the Markov transition matrix. Thereby, we have to find a way of working with a finite version, i.e., a truncation, of the original matrix.

4.2. Steady-State Analysis

We consider a subset of the state space for which $0 \le X(t) \le D_{\max}$. $D_{\max}$ is the maximum number of outstanding replenishment orders we consider. Since the second dimension of the state space, the number of class 2 backorders, is limited by the number of outstanding replenishment orders through (4.2) and (4.3), no truncation is needed for this dimension, i.e., $B(t) \le D_{\max} - (S-K)$. One should note that when we truncate the state-space as described, we are effectively ignoring an infinite number of states whose total probability is equal to $P\{X(t) > D_{\max}\}$. Since *X(t)* has a Poisson

distribution with parameter λL as stated in (4.4), we are ignoring the tail of Poisson distribution. Thereby, as $D_{\max}$ increases, the probability corresponding to the ignored part of the state space goes to zero rapidly.

In order to perform the steady-state analysis, we need to generate the Markov transition matrix for the chain. Since we are unable to generate the full matrix, we generate a submatrix of the transition matrix, which corresponds to the states that are conserved by the truncation, i.e. that are not ignored. We call this submatrix **Q**. The number of states corresponding to **Q** is $(S-K)+(D_{\max}-(S-K)+1)(D_{\max}-(S-K)+2)/2$. When $X(t)<(S-K)$, then $B(t)=0$, which means there are exactly $(S-K)$ states corresponding to that part of the state space. When $X(t)\geq(S-K)$, then the number of feasible $B(t)$ values increases one by one starting from one for $X(t)=(S-K)$. Then the total number of states in this part of the state space forms an arithmetic progression, which is $(D_{\max}-(S-K)+1)(D_{\max}-(S-K)+2)/2$ at $X(t)=D_{\max}$.

We consider a lexicographical ordering of these states in order to map the states to the columns and rows of the transition matrix. Then using the same idea one can easily map the state (x,b) to the $r(x,b)^{th}$ column and row in the transition matrix, where

$$r(x,b)=\begin{cases} x+1, & \text{if } x\leq S-K \\ x+1+(x-(S-K))(x-1-(S-K))/2+b, & \text{if } x>S-K. \end{cases} \tag{4.17}$$

The number of states increases with $D_{\max}$ and is in the order of $D_{\max}^2$. The recursive procedure discussed in the previous section computes the probability given in expression (4.8), for all feasible x_L, b_L, b, y, and z values.

Each of these dimensions is bounded by $D_{\max}$, since our truncation neglects the event that the number of demand or replenishment arrivals during leadtime is greater than $D_{\max}$. This means the computational complexity of the recursive algorithm is $O\left(D_{\max}^5\right)$. Thus, as a result of a computational effort of $O\left(D_{\max}^5\right)$, we are able to obtain the matrix **Q**, which is a finite submatrix of the original Markov transition matrix, which is infinite.

One could claim that once the truncation is performed, one does not have the original Markov chain, thereby an analysis based on this truncation would only be approximate. Although this claim is correct in the strictest sense of approximation, there is a theory in computational linear algebra, which states that one can get exact upper and lower bounds for steady-state probabilities corresponding to the nontruncated states by considering a truncated version of an irreducible Markov chain. The reader is referred for the theory behind the procedure giving the bounds to a paper by Courtois and Semal (1984).

It is observed that the upper and lower bounds for our problem converge together rapidly as $D_{\max}$ is increased. This is due to the additional structure our Markov chain possesses. Thus, we can get the steady-state probabilities with any desired accuracy.

In order to explain the procedure suggested by Courtois and Semal (1984), let us call the Markov transition matrix **P**. The partitioning of **P** into submatrices corresponding to truncation states and to ignored states can be expressed as

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{A} \\ \mathbf{B} & \mathbf{C} \end{bmatrix}. \tag{4.18}$$

One should note that **Q** is not a transition matrix, and thereby **Q** and (**I-Q**) are invertible given that **P** is a transition matrix corresponding to an irreducible chain. Let $\boldsymbol{\pi}$ be the steady-state probabilities vector corresponding to truncation states and $\boldsymbol{\pi}^{\mathbf{i}}$ the one corresponding to ignored states. Part of steady-state equations can be given as

$$\boldsymbol{\pi}\mathbf{Q} + \boldsymbol{\pi}^{\mathbf{i}}\mathbf{B} = \boldsymbol{\pi} \;\Rightarrow\; \boldsymbol{\pi}(\mathbf{I}-\mathbf{Q}) = \boldsymbol{\pi}^{\mathbf{i}}\mathbf{B} \;\Rightarrow\; \boldsymbol{\pi} = \boldsymbol{\pi}^{\mathbf{i}}\mathbf{B}(\mathbf{I}-\mathbf{Q})^{-1} \tag{4.19}$$

Equation (4.19) means that we could find steady-state probabilities corresponding to truncation states, if the vector $\boldsymbol{\pi}^{\mathbf{i}}\mathbf{B}$ were known. Based on this observation, one can construct another Markov transition matrix, $\tilde{\mathbf{P}}$, the truncated chain's transition matrix, by lumping all the ignored states into a single state, i.e.

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{Q} & \mathbf{1}-\mathbf{Q}\mathbf{1} \\ \mathbf{x} & 1-\mathbf{x}\mathbf{1} \end{bmatrix} \quad \textit{where } \mathbf{1} = \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^T \tag{4.20}$$

The row vector **x** represents the probability vector at which the process corresponding to the truncated chain makes a transition from the lumped state to the other states.

Let us consider the matrix $\tilde{\mathbf{P}}$ when $\mathbf{x} = (1/\boldsymbol{\pi}^{\mathbf{i}}\mathbf{1})\boldsymbol{\pi}^{\mathbf{i}}\mathbf{B}$. The vector $\boldsymbol{\pi}^{\mathbf{i}}\mathbf{B}$ can

be interpreted as the probability vector at which the process makes a transition from the ignored states into truncated states. Then $(1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1})\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}$ is the same transition probability under the condition that the process is in the ignored part of the state space. Therefore, the lumped state here mimics the dynamics of the ignored states. The steady-states probabilities corresponding to the matrix (4.20), will be the same with the ones of the original chain given that $\mathbf{x} = (1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1})\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}$. In order to show this, let us call the steady-state probabilities vector corresponding to truncation states under the new transition matrix $\tilde{\mathbf{P}}$ as $\tilde{\boldsymbol{\pi}}$ and the steady-state probability for the lumped state as π^{l}. Steady-state equations corresponding to truncation states are

$$\begin{aligned}\tilde{\boldsymbol{\pi}}\mathbf{Q} + \pi^{l}\left(1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}\right)\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B} = \tilde{\boldsymbol{\pi}} \;\Rightarrow\; \tilde{\boldsymbol{\pi}}\left(\mathbf{I}-\mathbf{Q}\right) &= \pi^{l}\left(1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}\right)\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B} \\ \Rightarrow\; \tilde{\boldsymbol{\pi}} &= \pi^{l}\left(1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}\right)\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}\left(\mathbf{I}-\mathbf{Q}\right)^{-1}\end{aligned} \tag{4.21}$$

The equation for the lumped state is not needed since it is linearly dependent with the equations in (4.21). This means that the steady-state probabilities can be obtained by setting π^{l} to 1, then finding the rest of the probabilities using (4.21), and finally normalizing them, i.e.

$$\begin{aligned}\tilde{\boldsymbol{\pi}} &= \frac{1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}}{1+\left(1/\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}\right)\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}\left(\mathbf{I}-\mathbf{Q}\right)^{-1}\mathbf{1}}\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}\left(\mathbf{I}-\mathbf{Q}\right)^{-1} \\ &= \frac{1}{\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}+\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}\left(\mathbf{I}-\mathbf{Q}\right)^{-1}\mathbf{1}}\boldsymbol{\pi}^{\mathrm{i}}\mathbf{B}\left(\mathbf{I}-\mathbf{Q}\right)^{-1} = \frac{1}{\boldsymbol{\pi}^{\mathrm{i}}\mathbf{1}+\boldsymbol{\pi}\mathbf{1}}\boldsymbol{\pi} = \boldsymbol{\pi}\end{aligned} \tag{4.22}$$

This means that the both transition matrices, $\mathbf{P}$ and $\tilde{\mathbf{P}}$, yield the same steady-state probabilities for truncation states. Thereby, one could obtain certain steady-state probabilities of an infinite state space Markov chain, using a

finite state space chain. The only problem is that in order to construct the finite chain, one needs to set vector **x** to $(1/\boldsymbol{\pi}^{i}\mathbf{1})\boldsymbol{\pi}^{i}\mathbf{B}$, which requires the steady-state solution of the original. Since the solution is what we are after, this result does not have practical relevance. Yet based on this result, exact upper and lower bounds can be developed for steady-state probabilities of interest.

The idea behind the procedure yielding the bounds is as follows: Since the vector $(1/\boldsymbol{\pi}^{i}\mathbf{1})\boldsymbol{\pi}^{i}\mathbf{B}$ representing the transitions from the lumped state to the truncation states in $\tilde{\mathbf{P}}$ cannot be determined without having the steady-state solution of the original chain, one can try to see what happens if any vector x of the same dimension is used instead of $(1/\boldsymbol{\pi}^{i}\mathbf{1})\boldsymbol{\pi}^{i}\mathbf{B}$. If we can find the steady-state probabilities vector solution set corresponding to all possible vectors, then the theory of Courtois and Semal (1984) states that the actual steady-state probabilities vector $\tilde{\boldsymbol{\pi}}$, which is equal to $\boldsymbol{\pi}$, has to be an element of this set. Moreover, it has been shown that this solution set forms a polyhedron, whose vertices are the solutions of the truncated chain with $\mathbf{x}=\mathbf{e}_{\mathbf{i}}^{T}$ for all i's where $\mathbf{e}_{\mathbf{i}}$'s are the standardized basis vectors ($\mathbf{e}_{\mathbf{i}}=\begin{pmatrix}0 & \dots & 0 & 1 & 0 & \dots & 0\end{pmatrix}^{T}$). Since $(1/\boldsymbol{\pi}^{i}\mathbf{1})\boldsymbol{\pi}^{i}\mathbf{B}$ is an element of the convex hull defined by the standardized basis vectors, it makes sense that the solution corresponding to $\mathbf{x}=(1/\boldsymbol{\pi}^{i}\mathbf{1})\boldsymbol{\pi}^{i}\mathbf{B}$ is an element of the convex hull defined by the solutions corresponding to the basis vectors.

One does not have to solve a different system of equations for every basis vector. In order to find all the solutions, it is enough to compute the inverse of (**I-Q**). Let $\tilde{\boldsymbol{\pi}}_{\mathbf{x}}$ be the solution of the truncated chain when the vector x is used. Then,

$$\tilde{\boldsymbol{\pi}}_{\mathbf{x}} = \pi^{l}\mathbf{x}\left(\mathbf{I}-\mathbf{Q}\right)^{-1} \quad \Rightarrow \tilde{\boldsymbol{\pi}}_{\mathbf{x}} = \frac{1}{1+\mathbf{x}\left(\mathbf{I}-\mathbf{Q}\right)^{-1}\mathbf{1}}\,\mathbf{x}\left(\mathbf{I}-\mathbf{Q}\right)^{-1} \tag{4.23}$$

using the same steps with the derivations given in (4.21) and (4.22). Basically $\tilde{\boldsymbol{\pi}}_{\mathbf{e_i}^T}$ can be obtained by taking the i^{th} row of $\left(\mathbf{I}-\mathbf{Q}\right)^{-1}$ and then applying normalization to it. The computational complexity of the inverse operation with *LU* factorization is $O(n^3)$ where n is the dimension of the matrix **Q**. Thence, since only one inverse operation is needed in order to apply our procedure, the computational complexity of the truncation algorithm is also $O(n^3)$, where n is the number of states that are not ignored in the truncation.

The bounds for individual steady-state probabilities can be obtained, once the polyhedron including the steady-state probability vector is known, by constructing a larger rectangular polyhedron covering the original polyhedron. The constructed polyhedron is defined by inequalities involving one dimension at a time. These bounds are given explicitly in Dayar and Stewart (1997). Let $z_{i,j}$ be the j$^{\text{th}}$ element of the vector $\tilde{\boldsymbol{\pi}}_{\mathbf{e_i}^T}$, which is the steady-state probability corresponding to the j^{th} state in Markov chain defined by truncated chain with $\mathbf{x} = \mathbf{e_i}^T$. Then the upper and the lower bounds for the steady-state probability of state j are

$$\xi_j^{\inf} = \max\left\{\min_i\left(z_{i,j}\right); 1-\sum_{k\neq j}\max_i\left(z_{i,k}\right)\right\} \tag{4.24}$$

$$\xi_j^{\sup} = \min\left\{\max_i\left(z_{i,j}\right); 1-\sum_{k\neq j}\min_i\left(z_{i,k}\right)\right\} \tag{4.25}$$

In our problem, the size of **Q** is determined by the maximum demand

considered during leadtime, $D_{\max}$. Since the dimension of **Q** increases with $D_{\max}$ and is in the order of $D_{\max}^2$. This means the algorithm providing bounds on the steady-state probabilities has a computational complexity of $O\left(D_{\max}^6\right)$. The recursive algorithm generating the matrix **Q**, has a computational complexity of $O\left(D_{\max}^5\right)$ as we have discussed before. Thus, the computational complexity of the whole procedure we suggest is $O\left(D_{\max}^6\right)$. This means that as $D_{\max}$ increases the computation times increase accordingly. Yet, as $D_{\max}$ increases, a greater portion of the system's dynamics is included in the matrix **Q**, which means the quality of the bounds provided by our algorithm increases, i.e., bounds become tighter. We know that, when we set $D_{\max}$, we are effectively ignoring those states whose total steady-state probability is

$$P\{X(t) > D_{\max}\} = \sum_{x=D_{\max}+1}^{\infty} e^{-\lambda L} \frac{(\lambda L)^x}{x!} \tag{4.26}$$

Since the Poisson probability mass function tends to zero rapidly beyond its mean value, one should select a $D_{\max}$ that is larger than the mean demand during leadtime, and increase it until the bounds are tight enough to yield the desired accuracy. One should note that in a setting where $(S-1,S)$ is applied, the value of $D_{\max}$ should not be very large, although it can be large enough to cause multiple outstanding orders. Under these circumstances, it would make sense to increase the lot size in order to decrease the replenishment traffic.

In order to clearly demonstrate the tradeoff between the computation times and the quality of bounds, we provide the results of a numerical experiment. The algorithm described in this paper is coded in MATLAB™ computing language, and is run on a PC with a Pentium III processor of 1 GHz clock speed. One should note that MATLAB™ is an interpreted language and the PC used is far from state of the art, which means that the computation times can be easily improved upon by transferring the code to a compiled language such as C or FORTRAN.

We consider an inventory system of the type described at the beginning of Section 4.2, with the following parameters: $S = 6$, $K = 3$, $\lambda_1 = 3$, $\lambda_2 = 2$, $L = 1$. In Table 4.1, we provide results of our algorithm for the described system. We start our experiment with $D_{\max} = 5$, since the expected total demand during leadtime is 5. The tightness of the bounds is measured by the maximum bound gap, which is $\max_j \left\{ \xi_j^{\sup} - \xi_j^{\min} \right\}$. In order to see how the performance measures of interest are affected, estimates for fill rates of the first and second classes (β_1, β_2) are also reported in Table 4.1. Since the exact steady-state probabilities are not known, but upper and lower bounds are, the steady-state probabilities are assumed to be the midpoint in the interval. The probabilities ignored by the truncation are assumed to be zero, i.e.

$$\hat{\beta}_1 = \sum_{x=0}^{\min\{S-1, D_{\max}\}} \sum_{b=(x-(S-1))^+}^{x-(S-K)} \hat{P}\{X = x, B = b\} \tag{4.27}$$

$$\hat{\beta}_2 = \sum_{x=0}^{\min\{S-K-1, D_{\max}\}} \hat{P}\{X = x, B = 0\} \tag{4.28}$$

where $\hat{P}\{X = x, B = b\} = \left(\xi^{\sup}_{r(x,b)} + \xi^{\inf}_{r(x,b)}\right)/2$.

One should note in (4.27) that if $D_{\max} < S-1$, then all the probabilities needed to compute the fill rate of class 1, will not be obtained by our truncation procedure. The same is true for the fill rate of class 2, when $D_{\max} < S-K-1$ as seen in (4.28). Thus, the smallest value one should consider for $D_{\max}$ is *S -1*.

One should also note that there is no need to calculate the fill rate using the steady-state probabilities obtained through our procedure. Since the steady-state distribution for the random variable *X(t)* is Poisson as stated in (4.4), we know that

$$\beta_2 = \sum_{x=0}^{S-K-1} P\{X = x, B = 0\} = \sum_{x=0}^{S-K-1} P\{X = x\} = \sum_{x=0}^{S-K-1} e^{-\lambda L} \frac{(\lambda L)^x}{x!} \quad (4.29)$$

Using (4.29), we can compute the exact fill rate for the second class, which turns out to be 0.1246 for the setting considered in Table 4.1. Thus, we see that our procedure yields an estimate for β_2, which converges to the theoretical value as expected. Moreover, our class 1 fill rate estimate agrees with the results we obtained from simulation experiments.

Table 4.1 Numerical Experiment (S = 6, K = 3, λ_1 = 3, λ_2 = 2, L = 1)

D_{max}	Computation Time (sec)	Maximum Bound Gap	$\hat{\beta}_1$	$\hat{\beta}_2$
5	<1	0.2954	0.6160	0.1856
6	<1	0.2047	0.7189	0.1604
7	<1	0.1254	0.7698	0.1373
8	<1	0.0680	0.7972	0.1216
9	1.5	0.0329	0.8216	0.1147
10	2.5	0.0144	0.8499	0.1145
11	4	0.0058	0.8774	0.1175
12	6	0.0021	0.8980	0.1207
13	9	7.43e-4	0.9104	0.1228
14	13	2.41e-4	0.9166	0.1239
15	18	7.35e-5	0.9193	0.1244
16	25	2.11e-5	0.9203	0.1245
17	35	5.77e-6	0.9207	0.1246
18	47	1.49e-6	0.9209	0.1246
19	61	3.68e-7	0.9209	0.1246
20	81	8.64e-8	0.9209	0.1246

It is clear that maximum bound gap decreases uniformly as $D_{\max}$ increases. As the bound gaps decrease, the error in our steady-state probability estimates also decreases, which in turn causes the performance measure estimates to converge to their theoretical values as expected. Our method provides the exact steady state distribution with desired precision

just like any other computational procedure, which can provide results valid up to a certain precision. The only approximation involved in our method is a numerical approximation similar to the ones used when computed an integral numerically. In numerical integration, the numerical error decreases as the step-size is reduced. This is exactly the kind of behavior observed as $D_{\max}$ is augmented in our procedure. In our experimental setting the average number of outstanding replenishment orders is 5, which is quite high for an inventory system. For other systems with smaller λL's, convergence occurs at smaller values of $D_{\max}$.

In order to give an idea about the effectiveness of the approximations proposed for the same system, we also calculated the fill rate estimates proposed by Dekker et al. (1998) and Deshpande et al. (2003). As we discussed in Chapter 3, both of these approximations yield the same results, which are $\hat{\beta}_1 = 0.8149$ and $\hat{\beta}_2 = 0.1246$. Their estimates for the class 2 fill rate are exact. Yet, the error involved in their class one fill rate is striking.

Chapter 5

RATIONING WITH CONTINUOUS REPLENISHMENT FLOW

Current level of information and computer technologies enables us to consider more elaborate policies. Data interchange is very fast and cheap in electronic environments. Although the analysis of these elaborate policies are difficult and mostly intractable, it is possible to estimate the steady state behavior of the system, even faster, with simulation in a reasonable amount of time with new computer systems. In this chapter we propose a dynamic rationing policy that makes use of the available information, specifically the number of outstanding orders and their ages. Since we are unable to provide a mathematical analysis, for this complex policy, we conduct a simulation study to quantify the gains. We assume a setting in which the inventory is replenished according to the continuous review (r,Q) policy. The leadtime is a constant, L, and there are two demand classes with Poisson arrival rates λ_1 and λ_2.

Typical policies, like the critical level policy, use the information about the inventory position and the on hand inventory level to make the replenishment and rationing decisions. In addition to these, we try to

incorporate the information that the outstanding orders carry to the decision mechanisms. However, it is not easy to define the threshold level as a function of the ages of outstanding orders. Instead, we consider a constant threshold level and decide to modify the on hand stock level dynamically by using the information that outstanding orders provide. For this purpose, we propose a dynamic rationing policy and call it Rationing With Continuous Replenishment Flow. Throughout the text, the initials *RCRF* will be used in place of this policy. *RCRF* incorporates the outstanding orders in to the on hand inventory as if they arrive continuously within the leadtime. The part of the order that is treated as if it has arrived is proportional to the ratio of the age of outstanding order and the leadtime. The whole order quantity completes its arrival at the end of the leadtime, just like in the original process. The on hand inventory level is modified by continuously adding the arrived parts of the outstanding orders. *RCRF* makes the rationing decision based on the modified on hand inventory instead of the real on hand level. If the modified inventory level is above *K* and we have stock at hand, class 2 demands are satisfied, otherwise backordered.

To define *RCRF*, for any time point *t* when a class 2 demand occurred and the on hand stock level is below K, let us define $a_i(t)$ as the age of i^{th} oldest outstanding order, $X(t)$ as the number of outstanding orders, $OH(t)$ as the on hand inventory level and $OH_M(t)$ as the modified inventory level. We have $0 \le a_i(t) \le L$.

We can model the continuous flow of the replenishment orders by considering the linear case, i.e. the part of the order that is treated as if it has arrived linearly increases with its age. In this case, the age ratio is $\frac{a_i(t)}{L}$ and

the quantity $Q\sum_{i=1}^{X(t)}\frac{a_i(t)}{L}$ is added to the on hand inventory level. However, the linear case directly uses the continuous flow assumption of the replenishment orders. We can increase the quality of our assumption, and so can decrease the gap between the unknown optimal policy and *RCRF*, by taking the powers of the age ratio, i.e. $\left(\frac{a_i(t)}{L}\right)^n$. We consider the powers greater than 1. Because, for all i we have $0 \le \frac{a_i(t)}{L} \le 1$, and in order to fine-tune the effect of continuous flow we should decrease the effect of the age ratio which is possible by taking the powers greater than 1. In this thesis we consider the integer values of n in order to make it possible to find the best values of n using a simulation based search procedure.

To make the effect of the power n clear, let us consider two different cases at the time of a class 2 demand arrival and assume $X(t) = 2$. For the first case, assume that $\frac{a_1(t)}{L} = 0.9$ and $\frac{a_2(t)}{L} = 0.1$. For the second case, assume that $\frac{a_1(t)}{L} = 0.6$ and $\frac{a_2(t)}{L} = 0.4$. If $n = 1$, we are indifferent between the two cases because the ratios add upto 1 for both of them. However, if $n = 2$, for the first case $\left(\frac{a_1(t)}{L}\right)^2 = 0.81$, $\left(\frac{a_2(t)}{L}\right)^2 = 0.01$ and they add upto 0.82. On the other hand, for the second case, $\left(\frac{a_1(t)}{L}\right)^2 = 0.36$, $\left(\frac{a_2(t)}{L}\right)^2 = 0.16$ and they add upto 0.52. Therefore the value of the information gained from the outstanding orders diminishes as we go from the oldest to the youngest order and this effect gets stronger as n increases. The

difference between $OH_M(t)$ and $OH(t)$ is mostly due to the outstanding orders that will arrive in the very near future.

Then the modified on hand inventory level can be expressed as

$$OH_M(t) = OH(t) + Q\sum_{i=1}^{X(t)}\left(\frac{a_i(t)}{L}\right)^n \text{ , } 0 \le a_i(t) \le L, \forall i \text{ and } n \in Z^+ \qquad (5.1)$$

If there is no outstanding order at time *t*, i.e. $X(t) = 0$, then $OH_M(t) = OH(t)$ and as in the critical level policy we compare the on hand stock level with the critical level to make the rationing decision. If $OH(t) > K$, then $OH_M(t)$ is surely greater than *K* and the arriving class 2 demand is satisfied.

It is important to note that in the calculation of $OH_M(t)$, the oldest outstanding orders, i.e. the ones that will arrive first, do not have any direct superiority or priority over the other outstanding orders. Because, we are interested in total potential of outstanding orders that will turn out to be physical stock units in the future. Naturally, the ratio $\frac{a_i(t)}{L}$ is close to 1 for the oldest replenishment orders and so they contribute more to $OH_M(t)$. Therefore, there is no need to make any prioritization of outstanding orders. Moreover, defining the power *n* strengthen the effect of the outstanding orders that will arrive in near future.

In order to be consistent with *RCRF*, clearing mechanism associated to this policy should also consider the modified on hand inventory instead of the real on hand inventory level. This is so because, when a replenishment order arrives future projection of the on hand stock is important to make the allocation decision of the order quantity between increasing the stock level

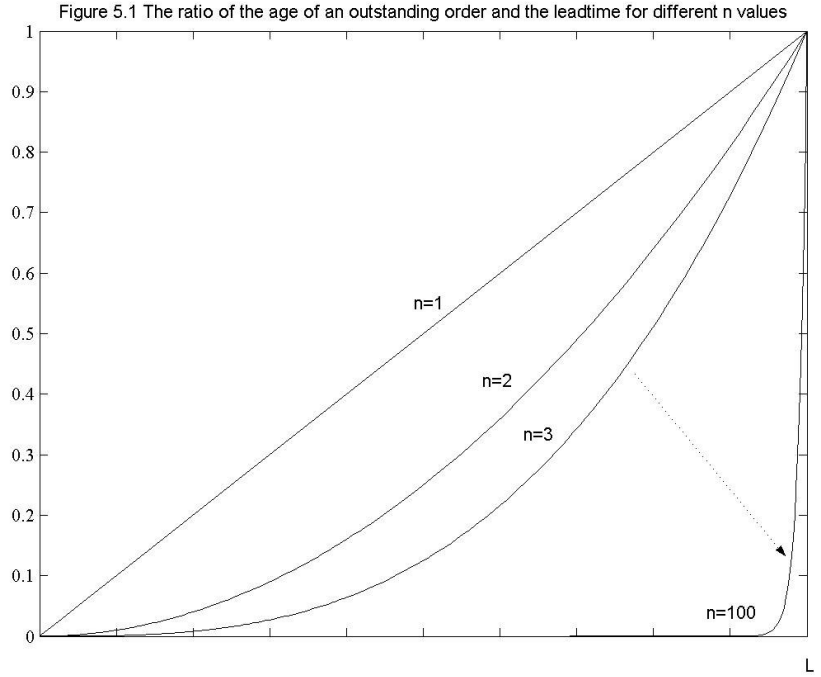
and clearing the backorders. Therefore, we use a modified priority clearing mechanism which compares the modified on hand stock level with K after clearing class 1 backorders. Suppose a replenishment order arrived at time t. Without taking the arrived order into account, we calculate $OH_M(t)$ using the remaining outstanding orders and current on hand level. If $OH_M(t) \geq K$ then we use the order quantity to clear class 2 backorders and add any remaining units to the stock. Otherwise, if $OH_M(t) < K$, if possible we increase $OH_M(t)$ up to K by adding the necessary amount of order quantity to the stock and then use the remaining order quantity to clear the class 2 backorders.

For given policy parameters, applying *RCRF* instead of the critical level policy decreases β_1 and increases β_2. Because, critical level policy stops serving class 2 when on hand stock drops below K, but dynamic policy provides service to class 2 customers whenever the modified on hand stock is above K even though the on hand stock is not. Similarly, the modified priority clearing mechanism of *RCRF* lowers γ_2, the average backorder time of a class 2 demand, compared to the priority clearing mechanism of critical level policy. Increasing value of β_2 contributes to the decrease in γ_2, because less number of class 2 customers is backordered so the total backorder time is less. In both clearing mechanisms upon arrival of replenishment order, clearing class 1 backorders has the highest priority. So modified clearing mechanism do not effect γ_1 directly. However, a lower β_1 value results in more class 1 backorders. Therefore we expect a higher γ_1 value under *RCRF*. To be able to increase β_2 and decrease γ_2 provides more incentive to the dynamic policy to increase K and this is what we observe from the simulation outputs. In most of the cases *RCRF* results in

higher K values than critical level policy.

An interesting feature of our proposed policy is it is possible to generate critical level policy from *RCRF*. As n goes to infinity, $\left(\frac{a_i(t)}{L}\right)^n$ goes to 0 for all i except for the replenishment orders that just arrived at time t for which the ratio is 1 and they are added to the on hand stock level at t. Then from (5.1), we have $\lim_{n\to\infty} OH_M(t) = OH(t)$. Therefore, rationing decision depends on the on hand stock level, i.e. as $n \to \infty$ *RCRF* turns out to be the critical level policy. For any replenishment order i, Figure 1 illustrates the power of the age ratio $\left(\frac{a_i(t)}{L}\right)^n$ for different n values.

Figure 5.1 The ratio of the age of an outstanding order and the leadtime for different n values



We start with $n = 1$ to illustrate the linear case, and as n increases the

value of the oldest outstanding orders increase. However, at the extreme, as $n \to \infty$, we no more use any information about the outstanding orders and do not make any dynamic decision. Our policy becomes identical with the critical level policy. Therefore, optimum n should be in the interval $(1,\infty)$. Except for the exponential leadtimes for which the age of outstanding orders provides no information, *RCRF* should perform better than the critical level policy. The optimum expected cost rate of critical level policy is an upper bound for the optimum expected cost rate of *RCRF*. In our case leadtime is constant, so for the given system parameters we can find (K,r,Q) and $n \in (1,\infty)$ values that give lower expected cost rate than the optimum of the critical level policy.

5.1 Performance Evaluation of *RCRF* with Simulation

To define how the gain through *RCRF* changes under different scenarios and to quantify the gain, we conducted a simulation study. In this study, for each parameter set we compare the minimum expected total cost rate values of *RCRF* and the critical level policy. Expected total cost rate, $E[TC]$, consists of expected values of holding, ordering and backordering costs as introduced in Chapter 3. Moreover, to get an idea about the benefit of rationing decision and to assess the relative value of dynamic stock decision we also simulate the common stock policy. We compare the optimal expected cost rates of the common stock policy and the critical level policy for each parameter set. As explained in Chapter 1, the common stock policy is one of the easiest ways of managing an inventory system that experiences different demand classes. In this policy both of the classes are served without any differentiation from a common stock pool. Rationing policies are alternatives of this policy in

order to provide different service levels to different customer classes.

Stock rationing is meaningful when $\pi_1 > \pi_2$ and $\hat{\pi}_1 > \hat{\pi}_2$, i.e. the unit backorder and the time-dependent backorder costs of class 1 are higher. To test the efficiency of *RCRF* under different pairs of backorder costs, we choose a low and a high value of π_1 and $\hat{\pi}_1$. We let π_1 take the values of 2 and 10; and let $\hat{\pi}_1$ 1 and 5. Then, π_2 and $\hat{\pi}_2$ are defined as the ratios of π_1 and $\hat{\pi}_1$. For each values of class 1 backorder cost, we set π_1 / π_2 and $\hat{\pi}_1 / \hat{\pi}_2$ to 5 and 1.25. Then if π_1 is 2, π_2 takes the values of 0.4 and 1.6. If π_1 is 10, π_2 can be 2 or 8. Similarly, if $\hat{\pi}_1 = 1$, $\hat{\pi}_2 = 0.2, 0.8$. And if $\hat{\pi}_1 = 5$, $\hat{\pi}_2 = 1, 4$.

The other most important factor that determines the performance of the rationing policies is the arrival rates of the customer demands. We fix the leadtime to 1, and change the arrival rates of both classes to generate different leadtime traffic rates. We let $\lambda = \lambda_1 + \lambda_2$ take values of 5 and 25, and set three levels of λ_1 / λ, 0.9, 0.5 and 0.1. Table 5.1 shows the arrival rate pairs for the simulation study.

TABLE 5.1 Arrival rates for the simulation study

	$\lambda = 5$	$\lambda = 25$
(λ_1, λ_2)	(4.5, 0.5)	(22.5, 2.5)
	(2.5, 2.5)	(12.5, 12.5)
	(0.5, 4.5)	(2.5, 22.5)

We set the holding cost to 5 and the fixed ordering cost to 2.

We define $E[TC_{RCRF}], E[TC_{CL}]$ and $E[TC_{CS}]$ as the expected total cost rates of *RCRF*, critical level policy and common stock policy correspondingly. Then we calculate the performance gain of *RCRF* as the cost reduction obtained by using the optimal *RCRF* policy compared to the optimal critical level policy:

$$G_{RCRF} = \frac{\min\{E[TC_{CL}] : r+Q > K \geq 0, Q \geq 1, r \geq 0\} - \min\{E[TC_{RCRF}] : r+Q > K \geq 0, Q \geq 1, r \geq 0\}}{\min\{E[TC_{CL}] : r+Q > K \geq 0, Q \geq 1, r \geq 0\}} x100.$$

Similarly, we compare the critical level policy with the common stock policy by calculating the percent gain

$$G_{RCRF} = \frac{\min\{E[TC_{CS}] : Q \geq 1, r \geq 0\} - \min\{E[TC_{CL}] : r+Q > K \geq 0, Q \geq 1, r \geq 0\}}{\min\{E[TC_{CS}] : Q \geq 1, r \geq 0\}} x100$$

Table 5.2 shows the performance gains, G_{RCRF} and G_{CL}, when $\lambda = 5$. For a specific (λ_1, λ_2) pair, the table displays maximum and minimum percent gains that observed for different backorder cost values. We also present the average value of the gain, i.e. average of the gains for different backorder cost values. The columns 2 to 4 present the gains through *RCRF* compared to critical level policy. And columns 6 to 8 show the gains through critical level policy compared to common stock policy.

As seen from Table 5.2, if we make the rationing decision based on *RCRF*

instead of the critical level policy, the maximum gain observed when $\lambda_1 = \lambda_2 = 2.5$. It is % 5.6. In this case, rationing is a valuable tool as we observe from G_{CL} values. On the average, the critical level policy provides %2.5 cost reduction and *RCRF* provides an additional %1.86 reduction. For both G_{RCRF} and G_{CL}, the maximum values are observed when $\pi_1 = 10$, $\pi_2 = 2$, $\hat{\pi}_1 = 1$ and $\hat{\pi}_2 = 0.2$. Therefore we can say that, when λ_1 and λ_2 are close to each other, if we compare *RCRF* and the critical level policy, *RCRF* provides most significant cost reductions when the gap between backorder costs of class 1 and class 2 is large. Similarly, under these conditions the critical level policy strictly dominates the common stock policy.

Table 5.2 Percent gain of *RCRF* over the critical level policy and percent gain of the critical level policy over common stock policy when $\lambda = 5$

	G_{RCRF}			G_{CL}		
(λ_1, λ_2)	Max.	Min.	Avg.	Max.	Min.	Avg.
(0.5, 4.5)	4.03	0.02	1.59	1.18	0.08	0.27
(2.5, 2.5)	5.16	0.04	1.86	7.29	0.08	2.50
(4.5, 0.5)	1.98	0.01	0.57	8.60	0.05	2.80

When we consider the case where the arrival rate of class 2 is larger than the

arrival rate of class 1, i.e. the case where $\lambda_2 = 4.5$ and $\lambda_1 = 0.5$, the critical level policy provides maximum %1.18 cost reduction. On the other hand, we get maximum % 4.03 and on the average % 1.59 cost reduction when we apply *RCRF* instead of the critical level policy. When the high portion of the total traffic is from class 2, it does not seem that rationing provides high savings if we just consider the critical level policy. However, our proposed dynamic policy provides considerable additional saving. Intuitively, critical level policy does not perform very well when λ_2 is very high compared to λ_1, because to eliminate large number of class 2 backorders K is set to 0. Therefore critical level rationing gets some gain only through the priority clearing mechanism. For the rationing problems it is hard to explain how the tradeoffs between holding, ordering and backorder costs affect optimum (K, r, Q) values when the input parameters are changed. However, we can say that *RCRF* can provide high savings when λ_2 is high, because, as we discussed before, *RCRF* decreases γ_2 and increases β_2. This fact provides much more gain when λ_2 is high. Thus compared to the critical level policy, *RCRF* does not result in high backordering costs for class 2.

For the same case, i.e. $\lambda_2 = 4.5$ and $\lambda_1 = 0.5$, maximum gain achieved through the critical level policy, %1.18, is observed when $\pi_1 = 10$, $\pi_2 = 8$ and $\hat{\pi}_1 = 5$, $\hat{\pi}_2 = 4$. Backordering costs of both classes are close to each other and so there is no distinctive difference between class 1 and class 2. This is in parallel with intuition because the critical level rationing sets K to 0 due to the high arrival rate of class 2 so there is not much need for critical level policy. Classes are not very distinguishable. Maximum value of G_{RCRF}, % 4.03, is observed when $\pi_1 = 2$, $\pi_2 = 1.6$ and $\hat{\pi}_1 = 5$, $\hat{\pi}_2 = 4$. Again

backordering costs of class 2 are at their maximum for the corresponding class 1 backordering costs.

If we analyze the case where $\lambda_1 = 4.5$ and $\lambda_2 = 0.5$, we observe that the critical level rationing provides significant cost reductions but *RCRF* cannot provide significant additional cost savings. Maximum value of G_{RCRF} is % 1.98. This can be due to the fact that when λ_1 is higher, γ_1 and β_1 are much more important. Since the expected number of backorders per unit time is $\lambda_1\beta_1$. For the same β_1 value, the case with a larger value of λ_1 results in more backordering cost than the case with a lower λ_1. For the optimum (K, r, Q) values of critical level policy, *RCRF* provides higher β_2 and lower β_1 values compared to critical level policy. Thus *RCRF* increases class 1 backorder costs much more than the decrease in the class 2 backorder costs due to the high λ_1. Therefore, the value of the information about the outstanding orders seems less significant in this setting.

We observe that when $\lambda = 5$, for all different values of the backorder costs, $n^* \in [4, 6]$ where n^* stands for the optimum value of the power n.

Table 5.3 shows the cost reductions for the case where $\lambda = 25$, in a similar way that Table 5.2 does for $\lambda = 5$.

All the comments that we made for Table 5.2 is also valid for Table 5.3, i.e. maximum values of G_{RCRF} and G_{CL} are observed when the arrival rates of class 1 and class 2 are equal and *RCRF* does not provide any significant additional cost saving when the arrival rate of class 1 is very large compared to the rate of class 2.

Moreover, when the maximum values of G_{RCRF} and G_{CL} that are obtained from Table 5.3 and from Table 5.2 are compared, the values from Table 5.3 are close to twice of the ones obtained from Table 5.2. This means that for the cases that the total traffic is high; applying *RCRF* instead of the critical level policy can result in more dramatic cost savings compared to the cases where the traffic is lower. Because, when the total demand rate increases, mostly the number of outstanding orders increases.

Table 5.3 Percent gain of *RCRF* over the critical level policy and percent gain of the critical level policy over common stock policy when $\lambda = 25$

	G_{RCRF}			G_{CL}		
(λ_1, λ_2)	Max.	Min.	Avg.	Max.	Min.	Avg.
(2.5, 22.5)	7.92	0.60	2.97	3.43	0.02	0.81
(12.5, 12.5)	10.70	0.33	4.49	13.48	0.12	5.03
(22.5, 2.5)	1.60	0.01	0.66	7.13	0.01	3.27

For the (r,Q) policy with backordering, expected number of outstanding orders is $\frac{\lambda L}{Q}$ (see Hadley and Whitin (1963) page 187). In our simulation study, when $\lambda = 5$ the maximum cost reduction achieved through *RCRF* when the optimum Q is 4, i.e. $Q^* = 4$. When $\lambda = 25$, $Q^* = 5$. Therefore

since L is fixed to 1, when we increase λ from 5 to 25, expected number of outstanding orders increases from 1.25 to 5 for the cases that we get maximum cost savings. Since *RCRF* uses the information about the outstanding orders, it gets more information when there are more outstanding orders and provides more cost savings.

When $\lambda = 25$, we observe that n^* cannot be defined in single short interval. If $\lambda_1 = 22.5$, $n^* \in [20, 22]$. If $\lambda_1 = 12.5$, $n^* \in [29, 33]$ and if $\lambda_1 = 2.5$ then $n^* \in [9, 15]$.

As a final observation, in Table 5.4 and Table 5.5, we present percentage of the cases where *RCRF* provides improvements for each component of the expected total cost rate function. For each parameter set, the comparison is made for the cases where the policies have their minimum expected total cost rates. Table 5.4 illustrates the percentage improvements for $\lambda = 5$, and Table 5.5 does for $\lambda = 25$. For both tables, columns 2 to 4 present the percentage of cases where β_1, β_2 and both β_1 and β_2 increases correspondingly. Similarly, columns 5 to 7 present the percentage of cases where γ_1, γ_2 and both γ_1 and γ_2 decreases. Column 8 illustrates the percentage of cases where *RCRF* decreases the average inventory compared to the critical level policy. Finally, column 9 presents the percentage of cases where *RCRF* decreases λ / Q, which determines the expected fixed ordering cost. For each parameter set, *RCRF* results in a lower expected cost rate than the critical level policy. Tables 5.4 and 5.5 illustrate how each cost component contributes to the improvement in the expected cost.

Table 5.4 Percentage of the cases where *RCRF* provides improvements in cost components when $\lambda = 5$

	$\lambda = 5$							
(λ_1, λ_2)	β_1	β_2	β_1, β_2	γ_1	γ_2	γ_1, γ_2	$\bar{I}$	λ/Q
(0.5, 4.5)	62.5	56.25	43.75	81.25	43.75	25	37.5	12.5
(2.5, 2.5)	50	56.25	37.5	62.5	18.75	12.5	43.75	18.75
(4.5, 0.5)	68.75	62.5	31.25	75	31.25	12.5	31.25	12.5

Table 5.5 Percentage of the cases where RCRF provides improvements in cost components when $\lambda = 25$

	$\lambda = 25$							
(λ_1, λ_2)	β_1	β_2	β_1, β_2	γ_1	γ_2	γ_1, γ_2	$\bar{I}$	λ/Q
(2.5, 22.5)	75	62.5	37.5	93.75	12.5	12.5	68.75	6.25
(12.5, 12.5)	75	93.75	68.75	93.75	18.75	18.75	56.25	25
(22.5,2.5)	87.5	75	62.5	93.75	62.5	56.25	18.75	37.50

Chapter 6

CONCLUSION

In this thesis, we consider the stock rationing policies for continuous review inventory systems. Firstly, we present a detailed analysis of the critical level policy and the backorder clearing mechanisms that used under the critical level policy. In this analysis, we position the works of Dekker et al. (1998) and Deshpande et al. (2003) by pointing out some ambiguities resulting from the literature.

Afterwards, we provide a new method for the analysis of continuous-review lot-per-lot inventory systems with backordering under rationing policy. Our method culminates in an algorithm to compute the exact steady-state distribution for the inventory system. Although there are many approximate results in the literature on this important inventory system, there was no method to obtain the exact steady-state distribution of the considered system up to this point. This constitutes the main contribution of the study.

The second contribution is the approach used in developing the algorithm. The method is based on the observation that continuous-review inventory systems with backordering evolve according to a Markov chain at multiples of its leadtime. This means that when the system is sampled at multiples of leadtime, a discrete-time Markov chain is obtained. Moreover, the steady-

state probabilities of this embedded Markov chain at hand are also valid for the underlying continuous-time inventory model. We believe that this new approach should be applicable to other continuous-review inventory systems with backordering as well. Therefore, the approach constitutes a contribution in its own right.

The third and final contribution relates to the computational procedure used in the algorithm for the computation of steady-state probabilities. Firstly, we provide a recursive algorithm to compute the probabilities defining the embedded chain. We note that since the state space of the embedded chain is infinite, one would need infinite amount of time to generate all the one-step transition probabilities, and then to obtain the steady-state probabilities. Then, we show that one can obtain upper and lower bounds for steady-state probabilities of certain states of a Markov chain using a truncated version of the chain. We explain how the quality of these bounds increase as the number of states conserved by the truncation is augmented. Finally, we show how the bounds can be used to obtain steady state probabilities of interest with desired accuracy. Although the theory behind the bounds is not new, its application to the analysis of infinite state-space inventory systems is. These bounds, which are known to be loose under general settings, work extremely well for the inventory system under consideration. We believe the technique of using a truncated Markov chain in order to find steady-state probabilities with desired accuracy can be applied successfully to other inventory systems as well. Furthermore, the exposition of the theory behind the technique in the study is self-contained. We believe that this exposition, which extracts a practical tool from the field of computational linear algebra, could be of use to operations researchers.

The scope of the method introduced can easily be extended. For example, it is straightforward to generalize the method to more than two priority classes. The state space dimension would have to increase by one for each additional class. We believe that a computer program that implements the algorithm for any number of priority classes would be a handy tool. The described tool could also have policy optimization capability.

A direct extension of the method would be the generalization of the current model from lot-per-lot policy to (r, Q) policy. This would certainly increase the application area of the resulting tool. Another interesting extension can be the addition of a deterministic demand leadtime to the model. The addition of demand leadtimes to a stock-rationing environment leads to many interesting issues that need to be investigated. An interesting issue that can be addressed is the relation between the demand leadtimes and the optimal rationing levels.

Finally, we propose a dynamic rationing policy that uses the available information about the number and the ages of the outstanding orders. Simulation study shows that dynamic policy provides considerable gains. With the current technology of the computer systems, it is easy to implement such kind of dynamic policies. An important extension can be the analytical analysis of the proposed policy.

Bibliography

Arslan, H., S.C. Graves, T. Roemer. 2005. A single product inventory model for multiple demand classes. Working Paper, Sawyer School of Management, Suffolk University, Boston, MA.

Atkins, D., K.K. Katircioglu. 1995. Managing inventory for multiple customers requiring different levels of service. Working Paper, 94-MSC-015, University of British Columbia, Vancouver, BC, Canada.

Cohen, M.A., P. Kleindorfer, H.L. Lee. 1989. Service constrained *(s, S)* inventory systems with priority demand classes and lost sales. *Management Science* **34** 482-499.

Courtois, P.-J., P. Semal. 1984.Bounds for the positive eigenvectors of nonnegative matrices and for their approximations by decomposition. *Journal of the Association for Computing Machinery* **31** 826-838.

Dayar, T., W.J. Stewart. 1997. Quasi lumpability, lower-bounding coupling matrices, and nearly completely decomposable Markov chains. *Siam J. Matrix Anal. Appl.* **18** 482-498.

Dekker, R., M.J. Kleijn, P.J. de Rooij. 1998. A spare parts stocking policy based on equipment criticality. *Int. J. Production Economics* **56-57** 69-77.

Dekker, R., R.M. Hill, M.j. Kleijn, R.H. Teunter. 2002. On the (*S* -1, *S*) lost sales inventory model with priority demand classes. *Naval Research Logistics* **49** 593-610.

Deshpande, V., M.A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Science* **49** 683-703.

Evans, R.V. 1968. Sales and restocking policies in a single inventory system. *Management Science* **14** 463-473.

Frank, K.C., R.Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Operations Research* **51** 993-1002.

Gayon, J.P, F.D. Vericourt, F. Karaesmen. 2005. Stock rationing in a multi class make-to-stock queue with information on the production status. Working Paper, Laboratoire Genie Industriel, 92295 Chatenay-Malabry Cedex, France

Ha, A.Y. 1997a. Inventory rationing in a make-to-stock production sytem with several demand classes and lost sales. *Management Science*. **43** 1093-1103.

Ha, A.Y. 1997b. Stock rationing policy for a make –to-stock production system with two priority classes and backordering. *Naval Research Logistics.* **43** 458-472.

Ha, A.Y. 2000. Stock Rationing in an $M/E_k/1$ make-to-stock queue. *Management Science.* **46** 77-87.

Hadley, G., T. Whitin. 1963. *Analysis of Inventory Systems*. Prentice-Hall, Englewood Cliffs, NJ.

Kaplan, A. 1969. Stock Rationing. *Management Science* **15** 260-267.

Kocaga, Y.L. 2004. Spare parts inventory management with dekivery lead times and rationing. M.S.Thesis, Industrial Engineering Department, Bilkent University, Turkey.

Melchiors, P., R. Dekker, M.J. Kleijn. 2000. Inventory rationing in an (*s*, *Q*) inventory model with lost sales and two demand classes. *Journal of Operational Research Society* **51** 111-122.

Melchiors, P. 2003. Restricted time remembering policies for the inventory rationing problem. *International Journal of Production Economics* **81** 461-468.

Moon, I., S. Kang. 1998. Rationing policies for some inventory systems. *Journal of Operational Research Society* **49** 509-518.

Nahmias, S., W. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Science* **27** 1236-1245.

Stidham, S. 1974. Stochastic clearing systems. *Stochastic and their applications* **2** 85-113.

Teunter, R.H., W.K.K. Haneveld. 1996. Reserving spare parts for critical demand. Technical Report, Graduate School/Research Institute System, SOM, University of Groningen.

Topkis, D.M. 1968. Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes. *Management Science* **15** 160-176.

Veinott, A.F. 1965. Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. *Operations Research* **13** 761-778.

Vericourt, F.D, F. Karaesmen, Y.Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Managemnet Science.* **48** 1486-1501

Zhao, H, V. Deshpande, J.K. Ryan. 2005. Inventory Sharing and rationing in decentralized dealer networks. *Management Science* **51** 531-547